DOCUMENT RESUME

ED 148 838 TM 006 723

AUTHOR	Reilly, Richard R.; And Others
TITLE	Expert Assessment of Experiential Learning--A CAEL Handbook.
INSTITUTION	Cooperative Assessment of Experiential Learning Project, Princeton, N.J.
SPONS AGENCY	Ford Foundation, New York, N.Y.; Fund for the Improvement of Postsecondary Education (DHEW), Washington, D.C.; Lilly Endowment, Inc., Indianapolis, Ind.
PUB DATE	77
NOTE	100p.; Contains occasional small print
AVAILABLE FROM	Cooperative Assessment of Experiential Learning, American City Building, Suite 403, Columbia, Maryland 21044 ($6.00)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Adults; Bias; College Students; Essays; *Evaluation Criteria; *Evaluation Methods; *Evaluators; *Guidelines; Higher Education; Informal Assessment; Interviews; *Learning Experience; Performance Tests; Reliability; Standards; *Student Evaluation; Validity; Work Experience; Writing Skills
IDENTIFIERS	*Experiential Learning; Portfolios

ABSTRACT

Principles and guidelines for the use of expert judgment of experiential learning are outlined. The report deals with a number of basic issues that apply to expert judgment, such as the role of the evaluator in defining criteria, and structuring the assessment procedure so that it will be reliable and valid. The importance of establishing objectively defined standards is stressed. Four methods of assessment are described: interviews, assessment of student products, performance tests, and assessment of written materials. A number of suggestions for the improvement of assessment and a discussion of problems to avoid are included. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

CAEL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED148838

TM006 723

cael

EXPERT ASSESSMENT OF EXPERIENTIAL LEARNING – A CAEL HANDBOOK

RICHARD REILLY
RUTH CHURCHILL
ARNOLD FLETCHER
MYRNA MILLER
JUDITH PENDERGRASS
JANE PORTER STUTZ
JOHN I.D. CLARK

COOPERATIVE ASSESSMENT OF EXPERIENTIAL LEARNING

CAEL STEERING COMMITTEE/BOARD OF TRUSTEES

Morris T. Keeton,*# CHAIRPERSON
Provost and Vice President
Antioch College

Richard J. Allen*
Director, Division of Arts & Sciences
The Johns Hopkins University Evening College

George Ayers*# SECRETARY effective 10/76
Vice President and Dean for Academic Affairs
Metropolitan State University

Barbara A. Barbato*#
Director of Contract Center for Individualized Study
Webster College

Neal R. Berte*#
President
Birmingham Southern College

James D. Brown, Jr.*#
President
Thomas A. Edison College

Arthur W. Chickering*#
Vice President for Policy Analysis and Evaluation
Empire State College

Hortense Dixon*
Vice President for Urban Programming
Texas Southern University

John S. Duley*#
Director, Field Experience Program
Justin Morrill College
Michigan State University

Nirelle J. Galson#
Assistant Director, Independent Study
Degree Program
Syracuse University

Sheila Gordon*#
Associate Dean for Cooperative Education
LaGuardia Community College

Cyril O. Houle*
Professor of Education
University of Chicago

Winton H. Manning*
Vice President, Development
Educational Testing Service

Larraine R. Matusak#
Dean, College of Alternative Programs
University of Evansville

Jules O. Pagano,*# VICE CHAIRPERSON
Dean, Program Development and Evaluation
Florida International University

Jean M. Pennington*
Director, Continuing Education for Women
Washington University

Jane S. Permaul#
Dean, Experimental Educational Programs
University of California at Los Angeles

David H. Provost*#
State University Dean
New Program Development and Evaluation
California State University and Colleges

Gilberto de los Santos*
Dean of Instructional Development
El Paso Community College

Robert F. Sexton#
Executive Director, Office for Experiential
Education
University of Kentucky

Peter P. Smith,*# SECRETARY prior to 10/76
President
Community College of Vermont

William G. Thomas*
Director, New Dimensions
Los Angeles Community Colleges

Urban G. Whitaker#
Dean of Undergraduate Studies
San Francisco State University

*Member of the CAEL Steering Committee that supervised the project during the period when the work reported in this volume was carried out.

#Member of the CAEL Steering Committee and Board of Trustees that took office October 12, 1976.

CAEL is an educational association of 250 institutions of higher education and other educational organizations. It was chartered by the Regents of the State of New York in September, 1976 for the purpose of fostering experiential learning and the valid and reliable assessment of its outcomes.

CAEL (Cooperative Assessment of Experiential Learning) started in March, 1974 as a research and development project involving Educational Testing Service and a group of colleges and universities. Funding was provided initially by the Carnegie Corporation of New York and subsequently by the Ford Foundation, the Lilly Endowment, and the Fund for the Improvement of Postsecondary Education.

Through June 30, 1977 inquiries concerning activities, publications, or membership should be addressed to CAEL, Educational Testing Service, Princeton, NJ 08540.

Following June 30, 1977 all correspondence should be addressed to CAEL, American City Building, Suite 403, Columbia, MD 21044.

The price of this publication is $6.00 per copy.

EXPERT ASSESSMENT OF EXPERIENTIAL LEARNING— A CAEL HANDBOOK

Richard R. Reilly, Educational Testing Service
Ruth Churchill, Antioch College
Arnold Fletcher, Thomas A. Edison College
Myrna Miller, Vermont State Colleges
Judith Pendergrass, Metropolitan State University
Jane Porter Stutz, Educational Testing Service
John L. D. Clark, Educational Testing Service

Cooperative Assessment of Experiential Learning

Copyright © 1977 by CAEL. All rights reserved.

FOREWORD

This *Handbook* is one of several major CAEL assessment reports based upon CAEL Working Papers produced in 1975. CAEL followed an unusual developmental plan in validating its initial work and preparing these reports. In order to insure that its developmental work was fully reviewed and tried out prior to final publication, the CAEL Working Papers were used experimentally in approximately 80 CAEL institutions during 1975-76. In a formal review procedure users provided critique and suggestions that the authors used in preparing this revision. Validation of these current materials also included a series of field studies that addressed important issues concerning reliability, validity, and the value of assessment to the student. Those validation studies are reported in *The CAEL Validation Report*, 1976; aspects of the findings especially pertinent to content are summarized briefly in Chapter III of this *Handbook*. The Preface summarizes the more important issues with which the authors tried to deal in their revisions and the major changes that distinguish this *Handbook* from the earlier version, originally published as Working Paper No. 10, *The Use of Expert Judgment in the Assessment of Experiential Learning*.

Warren W. Willingham
CAEL Project Director

CONTENTS

Preface v
Richard R. Reilly

I. Introduction 1
Richard R. Reilly

II. Methodological Problems and Issues Related to the Use of Expert Judgment in Assessment 6
Richard R. Reilly

III. The Interview and Related Procedures 25
Judith Pendergrass, Jane Porter Stutz, and *Richard R. Reilly*

IV. Product Assessment 48
Ruth Churchill

V. Performance Assessment 63
Arnold Fletcher and *John L. D. Clark*

VI. Assessment of Written Material 79
Myrna Miller

A Concluding Note 88
Richard R. Reilly

LIST OF TABLE AND FIGURES

Table 1. Comparison of General Steps Recommended for Assessment of Experiential Learning with Suggested Steps for Assessment of Sponsored Learning 5

Figure 1. General and Specific Definitions for Standards of Writing Competence .. 9

Figure 2. Example of a Two-way Table Used to Analyze Content Validity 13

Figure 3. Example of an Overly General Rating Scale 17

Figure 4. A Behaviorally-anchored Scale for Writing Competence 18

Figure 5. A Matrix of Assessment Steps 89

PREFACE

The individualized and often highly unique nature of experiential learning poses a number of problems for the faculty members and administrators charged with the assessment of that learning. Standardized tests, for example, are usually inadequate or inappropriate. Assessment must, therefore, be placed in the hands of an "expert," someone who has specialized knowledge and experience in a relevant discipline or field. This *Handbook* is an attempt to familiarize those involved in assessment with some of the basic principles and practices in the application of expert judgment. Though others may find this book useful, it is primarily intended for faculty members, administrators, and consultants involved in assessing the experientially-based learning of students who are seeking academic credit for that learning.

Responses to the working draft of this *Handbook* suggested an extremely diverse audience, ranging from educational philosophers to engineers. Consequently, an effort has been made to avoid overly technical language and wherever possible to illustrate and clarify points with concrete examples. Those familiar with the working draft of this document will note several changes. The introduction now contains a more extensive discussion of the problem of choosing an expert judge as well as a brief overview of some basic principles. The two separate chapters on interviewing have been combined into one chapter with a consequent reduction in redundancy and what is believed to be a better organization. A brief "Concluding Note" that reminds the reader of some general assessment principles has replaced the chapter on "Use and Evaluation."

It is hoped that this *Handbook* will be useful in three ways: First, practical methods and procedures are suggested that can help institutions make more effective use of expert judgment. Second, some examples of approaches to assessment are presented that would, with some modifications, be suitable for application at an institution where experiential learning must be assessed. Finally, it is hoped that the principles and methods suggested here will help institutions to evaluate the quality of their assessment on a continuing basis.

The writing of this *Handbook* was the joint effort of many people. Richard R. Reilly, an Educational Testing Service staff member through most of the project and currently at American Telephone and Telegraph Co., prepared Chapters I, II, and the Concluding Note, and contributed to Chapter III. Ruth Churchill of Antioch College wrote Chapter IV, and Arnold Fletcher of Thomas A. Edison College co-authored Chapter V with John L. D. Clark of ETS, who was on leave at Thomas A. Edison College when the working draft was written. Chapter VI was prepared by Myrna Miller, who was at Empire State College when the writing began and is now with Vermont State Colleges. Judith Pendergrass of Metropolitan State University was the senior author of Chapter III. Jane Porter Stutz of ETS contributed to Chapter III and coordinated communications and the compilation of the manuscript.

The following people reviewed drafts of the Working Paper or the revised draft and made many helpful suggestions:

Thomas Donlon, Educational Testing Service
Sheila Gordon, LaGuardia Community College
Joan Knapp, Educational Testing Service

Robert Ramos, American Telephone and Telegraph Company
Robert Sexton, University of Kentucky
Dennis Tippo, Warren Wilson College
In addition, faculty and administrators from many CAEL institutions completed a questionnaire on the Working Paper and provided useful recommendations.

Finally, appreciation is expressed to Lorraine Simon, who typed the early drafts and helped in countless other ways during the project.

Richard R. Reilly

I. Introduction

Richard R. Reilly

The assessment of learning invariably requires human judgment. This is clear enough in the case of a college instructor who must assess the learning of students taking a course. Perhaps less clear is the extensive reliance on judgment in the development of so-called "objective" multiple-choice achievement tests. In the first instance, the instructor may use direct observation of oral and written performance as a basis for grading students. In the second instance, judgment is involved in the selection of content to be covered and in the writing of items. In both of these examples the person making judgments is presumed to have some special knowledge or expertise in the relevant subject-matter field.

The assessment of experiential learning certainly is no different in demanding expert judgment. Whether the learning has been derived from a sponsored or nonsponsored experience, judgment will be necessary to evaluate the amount and quality of what has been learned. In many instances particularly when prior, nonsponsored learning is being assessed, assessors are forced to rely on only a limited amount of information. It is especially important, then, that the assessors, in addition to being "experts" in a particular substantive domain, also be competent judges. Experts who are not properly prepared to conduct assessment jeopardize the equity of assessment for the institution, and more importantly, for the individual student. Competent judges, on the other hand, while not guaranteeing perfect assessment, can ensure an equitable and consistent treatment of all students.

Overlap Between Different Methods

Four major assessment techniques are discussed in this *Handbook*—interviews, product assessment, performance assessment, and assessment of written material. These are not the only ways in which to assess experiential learning, but it is believed that these four methods comprise the most useful and relevant approaches for applying expert judgment. Though separate chapters are devoted to each method, it will become apparent to the reader that areas of overlap exist across methods. To some extent distinctions between methods are arbitrary and involve a matter of emphasis rather than a clearcut demarcation. This is perhaps most obvious in the case of product assessment and performance assessment where some of the methods and suggestions can be used interchangeably. The *emphases* are decidedly different, however. In product assessment there is greater concern with *what* is produced and less with *how* it is produced, whereas the reverse is true of performance assessment. The fact that there is overlap is not necessarily undesirable. It will be found that for many assessment problems a combination of two or more procedures may be much more effective than a single method. The principles of product and performance assessment can be applied in tandem in those situations where both the process and the end product must be considered to yield a valid assessment. Likewise, interviews may provide a useful adjunct to other assessment methods in many situations. It should be made clear that the separate treatment of

the four assessment methods in no way implies that each method should always be used in isolation. Indeed, it is the authors' intent and hope that users will recognize the advantages of applying more than one technique.

Purposes of This Handbook

This *Handbook* has four major objectives. First, it is hoped that readers can develop a better understanding of some of the general methodological problems and issues related to the use of expert judgment. Issues such as reliability and validity should be understood and considered before applying *any* assessment technique.

A second objective is to provide the reader with specific practical suggestions related to each of four common applications of expert judgment. The discussion of the problems and pitfalls of interviewing, product assessment, performance assessment, and the assessment of written materials should be helpful to those engaged in the application of these methods.

A third objective is to present the most important principles of sound assessment as a series of steps or checks to follow in conducting assessment. The general principles are presented later in this chapter and reoccur in modified form at the end of each chapter dealing with a specific type of assessment method.

A final objective is to broaden the reader's awareness of various possible approaches to assessment through the presentation of practical examples and prototype procedures. Some readers may find that an adaptation of a suggested procedure meets a local assessment need.

Choosing Assessors

Most of the content of this *Handbook* assumes that assessors have been identified and assigned. For many assessment problems, however, the task of choosing appropriate assessors is a critical and difficult one. Whitaker has distinguished six groups of potential assessors: (1) student learners in sponsored programs; (2) prior learners seeking credit for unsponsored learning; (3) faculty teachers; (4) nonfaculty teachers or supervisors in a sponsored learning program; (5) participant-observers such as clients or co-workers; (6) outcome observers. This last group includes "... persons who have not participated in the learning process in any way, but are called upon to participate in the assessment of learning outcomes".[1] Unlike the other groups, outcome observers have no association with the learning process that might make them convenient and natural choices for assessors. This will occur most often with the assessment of prior learning where a person with expertise in a specific area must be chosen to conduct an assessment.

Normally, choosing an outcome observer is a two-stage process involving *identification* and *selection*. In the identification stage two or more assessors are identified as having expertise in the subject-matter field for which credit is sought. At this point it becomes necessary to raise a very basic question. Who is an expert? In response to this question we offer a working definition: "An expert is an individual having special skill or knowledge derived from experience, education, or training."

---

[1]Urban G. Whitaker, "Assessors and Their Qualifications," in Morris T. Keeton and Associates, *Experiential Learning: Rationale, Characteristics, and Assessment*. San Francisco: Jossey-Bass, 1976, p. 193.

The definition purposely does not imply any formal academic "credentials" so that individuals who have expertise because of *their* experiential learning (e.g., an avocational pursuit) can be included.

How, then, is one to identify an expert? Specific types of evidence that might be used to determine whether an individual is an expert in a given subject or field include the following:

1. Recommendations from two or three other experts in the same or a closely related field.
2. Published works (e.g., journal or magazine articles, books, etc.) in a given subject.
3. Technical products or artistic works in a related area.
4. A directly related formal academic degree.
5. Other formal credentials such as a license.
6. Awards or honors for achievement in the field.
7. Instruction or teaching experience in the area.
8. Supervisory experience of individuals working in the same field.
9. Membership in organizations or societies with a focus closely related to the subject area.

Although there are really no hard and fast rules for identifying experts, no single one of the foregoing suggestions should be relied upon solely. As many independent sources of verification as possible should be obtained.

Assuming that several experts have been identified, usually one expert is selected[2] to conduct the assessment. What are some of the considerations that govern *which* expert is selected? Whitaker offers some useful criteria, which he refers to as "essential assessor characteristics." The first of these characteristics is *subject-matter expertise*. Although all of the potential assessors identified will have some subject-matter expertise, some may be stronger in certain areas than others. The strengths and weaknesses of the potential assessors should be judiciously considered in terms of the specific learning being claimed (e.g., theoretical knowledge vs. practical skills).

A second criterion involves the *psychometric expertise* of the assessor. Whitaker defines psychometric expertise as " . . . secondary knowledge of the assessment process that is sufficient to enable one to select and adapt techniques and instruments that others have developed and perfected."[3] Increasing the psychometric expertise of assessors is precisely what this *Handbook* is about. However, the experience and knowledge of the expert with regard to conducting assessment should be taken into account during the selection stage.

A third characteristic of effective assessors is what Whitaker refers to as *familiarity with case data*. Outcome observers as a group are typically not very familiar with all aspects of the case being assessed. In many cases, however, assessors will know something about the learner, the teacher (if any), the learning process, or other aspects of the case at hand. In general, such knowledge will help make the assessment more valid, and this should be considered in selecting assessors.

---

[2]Although it is always desirable to have multiple assessors, it is recognized that in most cases this is not possible because of costs or other practical considerations.

[3]Whitaker, pp. 198-199.

The final two criteria that should be considered are the *objectivity* and *motivation* of the assessor. The expert judge should be free of prejudice toward the learner, the teacher, and any other aspects of the learning process and, in addition, should be highly motivated to conduct assessment in the most thorough and equitable manner possible.

Not all assessors selected will possess all of these characteristics, of course. Indeed, if it were always possible to select assessors meeting all of these criteria there would be no need for a handbook like this one. However, careful identification and selection of expert judges can considerably simplify the task of preparing assessors.

Organization and Scope

The other chapters of the *Handbook* present a great deal of information, as well as materials, that might be adapted for use in assessment. Chapter II deals with some of the basic technical and methodological considerations common to all judgmental methods of assessment. Since the later chapters assume some familiarity with these general concepts, it is recommended that Chapter II be read before any of the chapters dealing with specific techniques.

Chapters III through VI discuss specific assessment techniques. Chapter III discusses the device most commonly used to assess experiential learning—the interview. The different types of interviews are described, potential problems are discussed, and a variety of practical suggestions are made for the improvement of interviewing and related techniques. Finally, the steps in developing a structured interview are presented and illustrated. Chapters IV and V discuss the methods of product and performance assessment. Some common problems are described and suggestions for improvement are offered. Both chapters detail model procedures for the conduct of assessment. Because of the heavy role that written material plays in experiential learning, a separate chapter has been included to discuss some of the relevant issues, problems, and methods. Chapter VI also contains a variety of useful suggestions that might be applied to assessment of written material. A "Concluding Note" summarizes six basic assessment steps and how they should be applied to the techniques presented in earlier chapters.

An Overview of Some General Principles

Because so much emphasis is placed on a set of basic assessment principles in the *Handbook*, it may be helpful for the reader to briefly consider the general steps recommended at the end of Chapter II. To illustrate the common elements of all assessment problems, and at the same time provide an overview of some basic principles, the six basic steps of assessment are presented together with six recommended steps for assessment of sponsored off-campus learning in Table 1.

The first step in the assessment of sponsored learning is *Program Definition*, which has no general counterpart, since for a large group of experiential learners, namely prior learners, no set program ever existed. The first general step, *Establish Criterion Standards*, does have a direct counterpart in *Specification of Learning Outcomes*. In both cases pinning down exactly *what* is to be assessed is critical for effective assessment.

Table 1

Comparison of General Steps Recommended for Assessment of Experiential Learning with Suggested Steps for Assessment of Sponsored Learning

General Steps for Conducting Assessment	Steps for Assessment of Sponsored Learning
	I. Program Definition
I. Establish Criterion Standards	II. Specification of Learning Outcomes
II. Select and Structure Assessment	III. Establish the Assessment Procedure
III. Plan the Administration of Assessment	
IV. Relate Judgments to Observations	IV. Formative Assessment
V. Record and Report	V. Summative Assessment
VI. Monitor Assessment	VI. Evaluation of Assessment

The next two general steps, *Select and Structure Assessment* and *Plan the Administration of Assessment*, are embodied in the step called, *Establish the Assessment Procedure*, for sponsored programs.

The fourth and fifth general steps, *Relate Judgments to Observations* and *Record and Report*, are specific desiderata for both *Formative and Summative Assessment*, which are the next two steps for sponsored programs. Relating judgments to observations simply means that assessors should tie their assessments as closely as possible to the evidence at hand and not be influenced by extraneous or irrelevant factors. Once a judgment has been made about the learning, the assessor should record the results of assessment for administrative purposes and provide feedback to the student. This last step is particularly important in the formative assessment phase in sponsored programs where diagnostic information can help the student learn more effectively.

The final general step, *Monitor Assessment*, is closely paralleled by *Evaluation of Assessment* in sponsored programs. Both steps are motivated by the recognition that assessment procedures can almost always be improved and should be checked for major flaws or weaknesses.[4]

The following chapters offer considerable amplication of these general points. It is hoped that the reader will recognize that effective assessment will share these principles regardless of the technique or method being applied.

[4]For further discussion of sponsored experiential learning and its assessment, see John S. Duley and Sheila Gordon, College-Sponsored Experiential Learning—A CAEL Handbook, Princeton, N.J.: CAEL, Educational Testing Service, 1977.

II. Methodological Problems and Issues Related To The Use Of Expert Judgment In Assessment

Richard R. Reilly

Although assessment of experiential learning is a relatively new endeavor, judgment has been studied in enough other contexts so that a reasonably well-defined set of general principles can be derived and applied. One conclusion of this research confirms what many of us suspected—that expert judges are often quite fallible. More importantly, however, it has also been shown that it is possible to correct or alleviate many of the errors that assessors make. Given the implications that such errors have for the student, it is important for administrators and assessors to be aware of some of the major problems and issues surrounding judgmental assessment.

Before considering the types of errors that judges typically make and what to do about them, we should first consider the *role* of the expert judge in assessment. In the previous chapter we addressed the questions, "Who is an expert?" and "How are experts chosen?" We now consider the question, "What does the expert judge do in assessment?" The assessment process has been conceptualized as involving six stages:[1] (1) identification; (2) articulation; (3) documentation; (4) measurement; (5) evaluation; and (6) transcripting. Once the learning acquired through experience has been *identified* and related to an institutional or programmatic goal (*articulated*), it must be *documented* in some way. After this has been done, expert judgment can be applied to *measure* the nature and extent of the learning and then *to evaluate* whether that learning meets certain standards. The final stage involves concisely describing the learning and recording it on a transcript (*transcripting*). In some cases, the expert plays a role in all six stages. An expert might be involved in identifying learning, for example, or even in transcripting the results of assessment. Clearly, however, the points at which specialized expertise is most essential are in the stages of measurement and evaluation. In many instances, learning will have been identified, articulated, and documented before the expert judge becomes involved in the assessment process. It is the task of the expert to measure the documented evidence of learning, and then evaluate this learning by comparing it with standards for academic credit.

FUNCTIONS OF THE EXPERT JUDGE

Although the role of the expert judge will vary depending upon the institutution, the student, and the type of learning to be assessed, the following seven functions are suggested.

Criterion Definition. Assessment implies standards. The expert, therefore, should play a role in the specific definition of the standards against which the evidence presented by the student is to be judged. These standards may be unique and

[1]Warren W. Willingham, "Critical Issues and Basic Requirements for Assessment."

highly specialized, but the assessor should be able to make them explicit and put them into objective terms.

Selecting. In some cases the assessment procedure may be predetermined by the institution or chosen by the student, as in the case of a product presented by a student for credit. In many cases, though, once the criteria have been specified, the expert will decide which type of assessment procedure is most appropriate for eliciting the most relevant sample of behavior, performance, or other evidence. The expert may, in fact, be able to take advantage of existing assessment procedures that could be adapted and structured for a student with relatively little revision.

Structuring. The degree of control over the structure of the assessment process that an individual assessor has will vary depending upon the situation. Structure may be imposed by others involved in the assessment process, for example. But in many situations the assessor can decide whether the assessment will be highly structured (e.g., an oral examination) or relatively unstructured (e.g., a loosely patterned interview).

Adapting. The expert assessor can have considerable impact on the validity of the overall judgments made by adapting the assessment technique to the experiences and needs of the individual. Adaptation does not necessarily imply a lack of structure, but rather the selection of the fairest and most relevant information for the demonstration of a student's particular learning or competence.

Observation. The expert must observe a student's behavior, performance, or products before reaching a decision. The observations are the critical stimuli to which the expert makes a response in the form of a judgment or assessment. The competence with which an expert can make observations is determined by some factors beyond the control of the immediate assessment process, such as experience and intelligence. Emphasis on the other functions of the assessor's role can go a long way toward enhancing the quality and relevance of the observations made. Focusing the assessor's attention on the specific behaviors or the most relevant aspects of performance can be done with clear, concise definitions, for example.

Recording. In instances where assessment is relatively complex or where judgment occurs sometime after observation, it is extremely important that some record of observations be kept. Observations may be recorded electronically, by the taking of notes, or only by the central nervous system of the expert.

Judging. The most critical function of the assessor's role is, of course, the act of judging or quantifying the student's learning. It is within this aspect of assessment that the expert must somehow consider what evidence has been presented, eliminate what is irrelevant, weigh what is relevant, and finally balance this against some standard of competence. The act of judging itself, though certainly crucial, does not take place in a vacuum. The other functions of the assessor's role in helping select the relevant evidence or sample of performance, in structuring the

assessment process, etc., are of equal importance in minimizing the errors that may be made in judgment.

These seven functions of the expert assessor's role may not all be carried out by the same individual, nor is it necessary that all seven be performed with each new assessment. For some common assessment situations the functions of criterion definition, selecting, and structuring may be performed only initially, since the results will have general application for all students seeking the same type of credit. For highly individualized and unique learning experiences it may be necessary for all seven functions to be performed on an ad hoc basis with each new assessment.

THE PROBLEM OF ESTABLISHING STANDARDS AND LEVELS OF COMPETENCE

Because assessment can be viewed as a comparison between the evidence presented by a student and some fixed standard, it is essential that the assessor have a clear understanding of what standards are being applied for a given assessment. In a measurement sense, the simplest type of assessment would be one in which the student's evidence is compared against a standard and a decision is made as to whether the evidence meets the standard or not. In many instances, however, the assessment process calls for a finer discrimination. A student may be awarded from zero to ten credits, for example. In such cases more than one standard is obviously needed, and, in fact, a continuum with anchoring standards at various points along a scale may be implied. Some of the most critical kinds of errors made in judgment have been shown to result from a lack of well-defined standards or criteria. Failure to provide specific standards can create at least two different kinds of problems. First, an assessor may have a very specific set of subjective standards that may not be shared by other assessors. This results in inconsistency in the assessment process since the final judgment made may depend more upon the assessor's unusual standards than upon the relevant evidence presented. A second possible response to a lack of clear standards is that the assessors may leave the standards vague. Under these conditions assessors would be more likely to "play it safe" and never really commit themselves strongly in making judgments. This too, results in a failure to discriminate.

In order for assessments to be fair and valid, the standards used by assessors should be defined explicity and in as much detail as possible. Where several levels or categories of competence are involved, experts in the subject-matter area should establish corresponding standards before an assessment is made. Ideally, there would be a consensus among experts as to the number of categories or levels and the standards which define those levels. For the assessment of common types of learning experiences, it may be possible to have predefined levels and standards. For the assessment of more unusual or new learning experiences, it might be necessary to set down objective standards on an ad hoc basis. A three-stage process is suggested. The expert would begin with a set of common standards written in terms that cut across all areas of learning and consequently are very general. These initial definitions could relate to the assessment framework of the institution and its philosophical orientation. In the second stage the expert, or experts, would translate these general definitions or standards into more specific

Figure 1
General and Specific Definitions for Standards of Writing Competence

General Definitions	Specific Definitions
Competence is unquestionably well above minimal requirements for credit.	Writing is clear and precise with excellent vocabulary and grammatical usage. Easily adapts language and style for different purposes.
Competence is minimally sufficient for credit.	Writing is adequate for freshman level, though vocabulary is somewhat limited and minor errors in grammar are sometimes made.
Competence is slightly below minimal requirements for credit.	Writing is sometimes unclear with words occasionally used inappropriately. Major grammatical errors are occasionally made.
Competence is far below minimal requirements for credit.	Writing often is incoherent and disjointed. Vocabulary is extremely limited, and major grammatical errors are often made.

and objective definitions related to the assessment area. The final stage would involve selecting objective indices or behaviors that relate to the experts' definitions of each level. If several experts are involved in setting standards, the objective indices or behaviors can be first generated independently by each expert and then compared before a consensus is reached. Figure 1 presents an example of the first two steps in this suggested process for the area of writing competence. The definitions on the left are very general and could be adapted to fit a particular assessment framework. The corresponding specific definitions on the right translate the general definitions into more objective terms related to writing competence.

For relatively common types of learning, such as writing competence, this definitional process would be necessary only once. After good standards have been established, they can be applied repeatedly to similar assessment problems.

In many instances, however, the learning claimed will be of a highly unique nature and consequently standards will have to be established on an ad hoc basis. The same steps can be performed even if only one expert judge is involved. Making standards explicit will, in fact, serve as a highly useful framework for structuring the assessment procedure itself. Well-defined standards serve as the cornerstone for valid and reliable assessments, and in cases where an assessment procedure is not predetermined or fixed, can serve as guidelines for developing and conducting the assessment procedure. There is, in addition, a further benefit to be derived from this process. From a student's point of view, feedback given in a framework of highly specific standards will be much more useful than feedback given in a very general framework. Thus, the educative benefit of assessment will be enhanced as a direct outcome of a process designed to yield more consistent and valid judgments about learning.

THE ISSUE OF VALIDITY

In general terms, the validity of an instrument is regarded as the extent to which a process or procedure is measuring what it purports to measure. In terms of assessment, validity can be thought of as the extent to which the assessment really measures the student's true learning experience as related to the criteria established. Researchers in the fields of education and psychology have distinguished five different approaches to evaluating the validity of an assessment procedure.

Predictive validity refers to the accuracy with which assessments can be used to predict later performance in a related area. Predictive validation is of fairly limited usefulness in the context of experiential learning. There may be situations, however, where an institution may wish to evaluate an assessment procedure, in part, by how well the assessments agree with later performance in a closely allied field. For example, an assessment procedure devised for awarding credit for a prerequisite course could be evaluated by comparing the assessments made with measures of actual performance in the later course.

A second method is called *concurrent validity* and refers to a procedure whereby a set of assessments is compared with a set of immediately available independent measures of the same learning. For example, an assessment procedure could be evaluated on an experimental basis by comparing the average assessment made for a group of students with a known and well-documented level of competence in some specific area with the average assessment made for another group without such expertise. Such an approach could be used to evaluate the usefulness of a performance test designed to measure laboratory skills in chemistry. The average level of performance, as judged by experts, could be compared for graduate students in chemistry and undergraduates without extensive training. A large difference between the two groups in average performance would "validate" the assessment procedure.

Both predictive and concurrent validations are limited in usefulness because of the requirement that *groups* of students be assessed for similar sets of competences. The more individualized the assessment for each student, the more difficult it becomes to apply these empirical methods.

A third approach to the validation of assessment procedures is referred to as *construct validity*. The construct validity of an assessment procedure can be

thought of as all the accumulated supporting evidence for the usefulness of that procedure. Construct validity is a long-range process that evaluates the theoretical underpinnings of an assessment device or procedure by examining the relationships between the assessments and other measures which are thought to measure the same or similar sets of factors. Construct validation is a conceptual process used primarily to evaluate instruments that purport to measure traits such as intelligence or personality factors. The program of research implied by the concept of construct validation would probably apply in an active sense to only a very limited number of institutions with very special capabilities and interests. Construct validation comes into consideration in a different way for a larger number of institutions. In choosing an assessment procedure, or in setting up an assessment program, the construct validity of various techniques and procedures used would be an important consideration. For instance, in deciding whether or not the leaderless group discussion technique should be used to assess managerial skills, the available evidence relating to the construct validity of the leaderless group discussion for that purpose should be examined.

Content validity is the fourth approach to validation and the most useful in the present context. As a process, content validity involves a systematic examination of the assessment procedure to determine whether it is designed to elicit behaviors or indices that are relevant to the explicit standards established. Content validation is almost always a judgmental process and should, if possible, be done by someone other than the expert responsible for devising the assessment procedure. Ideally, several experts would evaluate the procedure from the point of view of representativeness and comprehensiveness. One approach to the content validation of an assessment procedure that may be helpful is first to develop a list of the specific objectives of an assessment. The content validity of a proposed assessment procedure can then be evaluated by considering whether or not each of the listed objectives is met.

A more elaborate version of the same idea would involve also breaking up the assessment procedure into modular components. The assessment of writing competence, for example, might be assessed through three separate documents: (1) a log in which the student has recorded learning experiences resulting from a sponsored work program, (2) an essay treating a general topic broadly related to the work experience, and (3) a report dealing with the history of an industry or profession closely related to the work experience.

A potential list of assessment objectives might include the evalulation of the following specific areas:

- vocabulary
- grammar
- ability to compare and contrast
- narration
- dialogue
- use of footnotes

A two-way configuration, as shown in Figure 2, could then be used to evaluate the content validity of the entire procedure. The entry in each cell could be a rating,

reflecting *how well* a given assessment component covers a given assessment objective, or simply a check indicating whether or not a particular component covers a certain objective. Use of this more elaborate procedure could be helpful in two ways. First, breaking up the assessment procedure focuses attention on the actual content of assessment and helps evaluators to avoid making unwarranted assumptions or inferences. Second, the ratings (or checks) can be summed both vertically and horizontally. The vertical sums indicate how well each objective is covered by the total assessment procedure, and the horizontal sums indicate the overall usefulness of each assessment component. Such information would be particularly helpful in revising an assessment procedure. If certain critical objectives were not being met, the assessment procedure could be extended to include measurement of those objectives. On the other hand, some components of the current assessment procedure might be shown as adding very little to the measurement of objectives and could thus be eliminated.

The example given in Figure 2 shows that competence in writing dialogue is not covered by any of the three methods. At this point a decision might be made to include a fourth writing sample, such as a short story, which would allow the student to demonstrate competence at writing dialogue. Breaking down the assessment process in this way illustrates how the method of content validation can be used to evaluate and revise an assessment procedure before it is even used. If certain critical objectives were not being covered by the assessment procedure, it could be changed or extended.

There is one additional type of validity often discussed by measurement experts. This is *face validity*, or the extent to which the assessment procedure appears, on its face, to be measuring what it is supposed to measure. For example, multiple-choice tests have been used to measure writing competence with considerable success. The *face validity* of a multiple-choice test is not especially good, however, because the task of choosing multiple-choice options does not appear to be as highly related to writing competence as, say, writing an essay. In this case the essay would have much better face validity than the multiple-choice test. As a general rule-of-thumb for applied measurement, face validity is always desirable. Assessment procedures lacking face validity will inevitably be questioned by students, especially those who do poorly. Assessments with high face validity tend to create a more positive atmosphere, which may, in fact, result in more valid assessments.

Any assessment procedure is only a sample of what a student has accomplished or can accomplish. In order to ensure validity, it is the task of the assessor to make this sample as relevant and as representative as possible. In some cases the assessor may be limited as to degree of control over the sample of performance or behavior. A product may be presented, for example, as an indication of a student's competence in some artistic or technical area. Even in this type of case, however, assessment can be made more valid by taking into account the context in which the product was created. Such assessments often use interviews as a way of gaining information about the context, and it is here that the assessor can lend some structure to the process to ensure representativeness and relevance.

Figure 2
Example of a Two-way Table Used to Analyze Content Validity

	Assessment Objectives						
	Vocabulary	Grammar	Ability to Compare & Contrast	Narration	Dialogue	Footnotes	Overall Usefulness of Each Component
Log	2	2	1	3	0	0	8
Essay	3	3	3	0	0	0	9
Report	3	3	2	0	0	3	11
Adequacy of Content for Each Objective	8	8	6	3	0	3	

Assessment Components

Rating Scale:

3 Should provide information highly related to assessment objective.
2 Should provide some information related to assessment objective.
1 Should provide very limited information related to assessment objective.
0 Should provide no information related to assessment objective.

RELIABILITY OF ASSESSMENT

If validity describes how well an assessment procedure measures what it is supposed to measure, then reliability can be said to describe how consistently an assessment procedure measures what it is measuring.

Intuitively, the notion that judgments should be made on a consistent and reliable basis seems necessary from the point of view of fairness to the student. If the same individual is given widely varying assessments depending upon the judge or the assessment technique used, one would surely question the basic equity of the assessment system. The implications that reliability of judgment has for the validity of judgment may not be as obvious, however. As a general rule, judgments can be consistent without being valid, but judgments can never be valid without being consistent. Reliability is, in other words, a necessary but not sufficient condition for validity.

What Factors Affect Reliability of Assessments?

It is important to recognize three major components in the assessment process, each of which can contribute to the unreliability of judgment: the task, the student, and the assessor. First, there is the "task" or assessment situation itself. A good illustration of an unreliable task is an oral test with only two or three questions. Although it is unlikely that such a test would be used in practice, the probability of getting an accurate measure of competence from such a test is extremely low since the assessment result would probably depend heavily on chance factors related to whether or not the questions happen to hit areas with which the student is familiar.

The second major component of the assessment process that can contribute to the error in the overall assessment is the student, or person being assessed. Despite their lack of relevance to actual knowledge or competence, short-term changes in mood, physical fatigue, and shifts in anxiety level are some of the factors that could cause students' performance to differ.

The final component is, of course, the assessor. The assessor brings a set of idiosyncratic standards, attitudes, and preferences that can lead to errors in assessment. For example, a rater with extremely harsh standards will make a different assessment from a rater with very lenient standards. It is this last component that concerns us most directly. From the research that has been done on the consistency and accuracy with which assessors make judgments, a variety of specific sources of error have been identified. Some of the more important sources of error are as follows:

1. Leniency or Harshness Error. Some assessors tend to make judgments that are, on the average, much more favorable or more lenient than judgments made by other assessors. Conversely, other assessors may make judgments that are consistently more unfavorable than the judgments of other assessors. These errors are sometimes referred to as leniency and harshness effects and are analogous to the phenomenon that students in traditional settings encounter in "easy" or "hard" graders.

Example. Over a period of six months two professors both evaluate indepen-

dent reports submitted by 30 students seeking academic credit for a prior learning experience. Professor A grants credit to 27 of the students, but Professor B grants credit to only 8 students. The assessments are inconsistent because Professor A clearly has more lenient standards than Professor B.

2. Errors of Central Tendency. Many assessors are reluctant to commit themselves one way or the other and as a consequence tend to make most ratings near the average or center of a scale. This type of error, often referred to as the error of central tendency, is particularly troublesome in situations where some discrimination among individuals in a group is needed. In the individual assessment situation, errors of central tendency will result in a lack of discrimination and lower reliability and validity.

Example. A panel assessment procedure involving four judges is used to evaluate the competence of five different students at piano. Three of the judges agree perfectly and rate two of the students "below standard," two "outstanding," and one "adequate." The fourth judge, who is less experienced, "played it safe" and rated all five students "adequate." In doing so, he failed to discriminate among levels of student performance.

3. Halo Effect. In situations where a student is being assessed in several different specific areas, a favorable overall impression may result in unjustifiably favorable judgments in all areas. This type of error is often referred to as the "halo" effect although the reverse can also occur. That is, an unfavorable overall impression can result in unjustifiably unfavorable judgments in specific areas.

Example. A highly articulate, personable student presents three short stories for academic credit in English. Two professors evaluate the stories. Professor A has known the student informally for about a year and considers her highly intelligent. Professor B has never met the student. Professor A judges the stories to be of acceptable quality for credit, but Professor B does not. One possible explanation for the disagreement is that the halo effect caused Professor A to rate the stories too high.

4. Initial Impressions. An error similar to the halo effect can result from the initial impression an assessor has of a student. In some situations it has been shown that a favorable or unfavorable initial impression will unjustifiably affect later judgments of specific aspects of performance.

Example. A student about to be interviewed about his travel experiences ignores the assessor's outstretched hand and takes a seat. The interview is conducted, and the assessor decides not to award credit. The unfavorable initial impression made by the student may have caused the assessor to adopt harder standards than would usually be the case.

5. Stereotypes. Strongly held attitudes or beliefs can cause misperception and error in judgment. A good example is the judge who is influenced by a stereotype of members of a particular class or group. It should be pointed out that such errors may be favorable or unfavorable to the student depending on the kind of stereotype held.

Example. A 65-year-old woman and a 25-year-old man both submit reports based on similar work experience. Though the reports are equivalent in quality, the man is awarded more credit The difference in judgments might have been caused by stereotypes held regarding older women.

6. Contrast Effect. The quality of the student who was rated previously will often affect judgment. An average student may tend to receive lower than average ratings if the previous student was outstanding and higher than average ratings if the previous student was poor. This type of error has been referred to as the contrast effect.

Example. An assessor interviews two students successively regarding similar work experiences. The first student gives extremely evasive, vague answers about the experience and seems generally unable to articulate any learning outcomes. The second student describes the experience and states learning adequately. The first student is judged to be "well below standard," but the second student is judged "outstanding" even though on an absolute basis the interview performance was only adequate.

7. Similarity of Background. The degree of similarity between a judge and the person being assessed with respect to background, attitudes, and ethnic group has been shown to affect judgments, with greater similarity tending to produce more favorable judgments.

Example. Two students, one black and one white, are both interviewed by one black and one white assessor. The black assessor rates the black student "outstanding" and the white student "average." The white assessor rates the black student "average" and the white student "outstanding." The discrepancy might be explained by the similarity or difference between assessors and students.

In addition to these specific classes of errors, other factors can affect the consistency and accuracy of judgment. It has been shown empirically and demonstrated theoretically, for example, that the average of several assessors' ratings is more reliable than the ratings of one assessor. This seems to make sense intuitively when the various sources of error described above are considered. Most of these are idiosyncratic or unique for a given assessor. The advantage of having several assessors or judges is that many of these errors will cancel each other out and thus produce more accurate or reliable ratings.

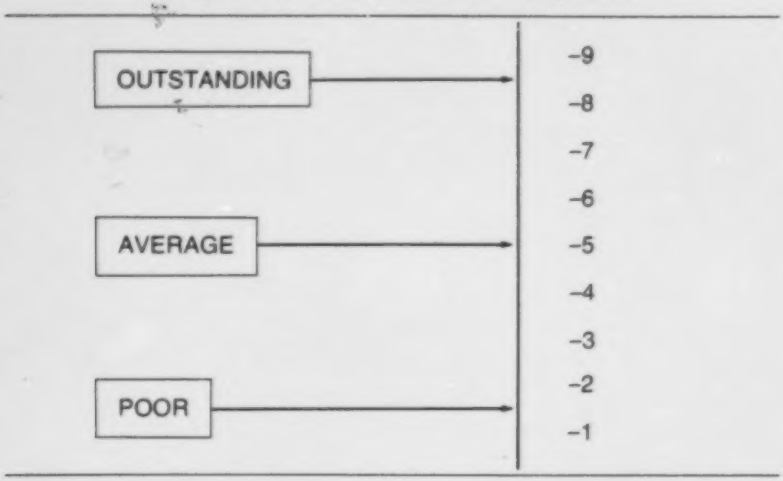
Another source of error may be found in the transition from observation to judgment, particularly if no systematic recording system has been used. Some studies have shown, for example, that an assessor's ability to recall specific information from an unstructured interview is often quite poor. Judgments may as a consequence be based on erroneous or irrelevant information.

Reducing Error

Error in an assessment process can never be completely eliminated, but research has shown that errors can be reduced. Most of the errors described above become

less serious as criterion standards become more explicit and objectively defined. The explicit definitions help to focus the assessor's attention on the specific competence being assessed and help to reduce conjecture and inference that may be based on irrelevant factors. A technique that can go a long way toward improving assessment, if used correctly, is to provide assessors with devices to aid observation, such as rating scales or checklists. Rating scales, if properly constructed, can serve both as a translation of criterion standards into objective behavioral terms and as a systematic aid in recording observations. A rating scale will only be worthwhile, however, to the extent that it helps clarify the criterion standards and focuses the attention of the assessor on specific, relevant evidence.

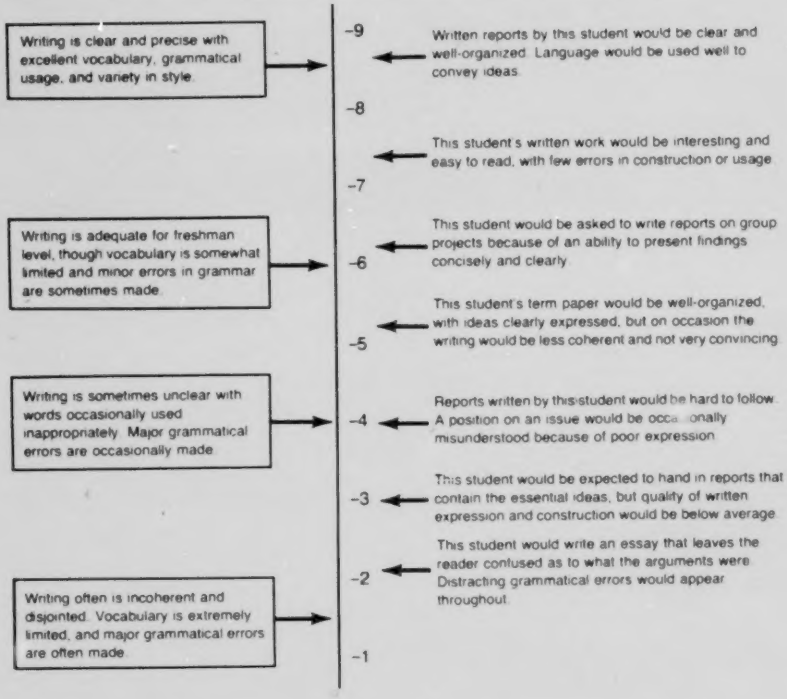
Figure 3
Example of an Overly General Rating Scale



A very general rating scale such as that presented in Figure 3 will clearly not accomplish these purposes. The terms "outstanding," "average," and "poor" are too general and unrelated to criterion standards to be of much help. A more useful approach is presented in Figure 4. Here the points of the scale, particularly the critical points, have been anchored by clear verbal descriptions related to the criterion standards described earlier in Figure 1. In addition, behavioral examples help to define the scale more fully. Rating scales of this type are referred to as behaviorally-anchored rating scales and can be developed in situations where criterion standards are explicitly defined.[2] The general definitions on the right hand side of this scale should follow directly from the established criterion standards. Specific behaviors should be generated and scaled by experts in the relevant subject-matter area. Behaviorally-anchored rating scales can be quite useful where relatively complex performance is being observed, as in a group discussion.

[2]A good description of how to construct behaviorally-anchored rating scales can be found in P. C. Smith and L. M. Kendall, "Retranslation of Expectations," pp. 149-155.

Figure 4
A Behaviorally-anchored Scale for Writing Competence


Writing is clear and precise with excellent vocabulary, grammatical usage, and variety in style.
Writing is adequate for freshman level, though vocabulary is somewhat limited and minor errors in grammar are sometimes made.
Writing is sometimes unclear with words occasionally used inappropriately. Major grammatical errors are occasionally made.
Writing often is incoherent and disjointed. Vocabulary is extremely limited, and major grammatical errors are often made.
-9
Written reports by this student would be clear and well-organized. Language would be used well to convey ideas.
-8
This student's written work would be interesting and easy to read, with few errors in construction or usage.
-7
This student would be asked to write reports on group projects because of an ability to present findings concisely and clearly.
-6
This student's term paper would be well-organized, with ideas clearly expressed, but on occasion the writing would be less coherent and not very convincing.
-5
Reports written by this student would be hard to follow. A position on an issue would be occasionally misunderstood because of poor expression.
-4
This student would be expected to hand in reports that contain the essential ideas, but quality of written expression and construction would be below average.
-3
This student would write an essay that leaves the reader confused as to what the arguments were. Distracting grammatical errors would appear throughout.
-2
-1

In other situations it would not be feasible or sensible to use behavioral examples (e.g., in assessing a product). In such cases the verbal descriptions may be all that is feasible or necessary. The effort involved in developing good rating scales is considerable and hence is worthwhile only for assessment situations that come up repeatedly.

In highly unique assessment situations it would not be justifiable to develop an actual rating scale, but the definitions that correspond to the criterion standards should be explicity considered when an assessor makes a judgment. The case of highly unique assessments illustrates one facet of a problem which assessing institutions or agencies must confront periodically. Although steps can be taken to reduce the error in any assessment procedure, this cannot be done without some cost. In the case of highly unique assessments, the expense of developing a rating system and a structured assessment procedure may be prohibitive.

No matter how carefully developed a rating scale is, however, it is important to recognize the importance of preparing or training assessors. Since many of the problems with judgmental assessment stem from individual assessors operating with different standards, a goal of any institution wishing to improve its assessment

should be to ensure that assessors are operating within a common framework of values and criteria.

EVALUATION OF ASSESSMENT

The periodic evaluation of assessment should be a permanent feature of any assessment program, but evaluation of assessment procedures should also be made whenever a new procedure is developed or revised. In all cases a staff member or consultant with some expertise in educational and psychological measurement should monitor the evaluation. If this is not possible, books such as those referenced at the end of this chapter should be consulted. Following are examples of some approaches that might be taken to the estimation of reliability and validity of assessment.

Assessing Reliability of Assessment

1. Reliability of Overall Assessment Procedure. As stated earlier, inconsistency or error in the assessment process can arise from the assessor, the student, or the assessment task or situation. Before attempting to separate these separate souces of error, an evaluation should be made of the reliability of the overall assessment procedure. If reliability is adequate, it may not be necessary to investigate sources of error.

The key to reliability estimation is in obtaining two or more independent assessments for a sample of students. To be reliable the extent of agreement among independent assessments should be high. Obtaining independent assessments might be relatively easy for products or written material, but can be quite expensive for interviews or performance assessment. An interview or performance would have to be repeated with a different expert or set of experts to obtain independent judgments. Although difficult, it is usually possible to carry out such procedures on an experimental basis. A random sample of students could be interviewed twice by different assessors, for example. It is important to keep in mind that the *entire* assessment procedure must be repeated independently in order to obtain an accurate estimate of assessment reliability. A product, for example, might be assessed by first interviewing a student for background information and then making a judgment. Both the interview and the judgment should be done independently by a second expert if reliability of the entire procedure is to be estimated.

2. Reliability of Judges. In order to estimate the reliability of each potential source of error, it is necessary to hold the other sources constant in some way. Thus, the reliability of judges can be assessed only if the sources of error due to students and the task can be controlled. Reliability of judgments obtained from interviews can be estimated by having two or more judges present during the same interview, or by having different judges make assessments of the same set of taped or videotaped interviews. Similar procedures could be used for performance assessment. For products and written material, a set of sample materials could be evaluated by different judges. Where interviews are employed as an adjunct, transcripts or recordings of the interviews could be made available to the judges.

3. Task Reliability. In situations where a structured task is involved, it is often possible to estimate the reliability of the task, apart from other sources of error. With some assessment procedures it may be possible to separate a given performance, interview, or simulation task into "parallel" parts—parts of equal difficulty that measure approximately the same things. A high degree of agreement between scores or judgments formed on the basis of performance of the two parallel parts would indicate good internal consistency of the assessment procedure. This might be done, for example, with a performance test consisting of two separately timed parallel parts.

4. Reliability of Student Performance. Estimating reliability of student performance is perhaps the most difficult source of error to isolate since it requires obtaining samples of performance, products, or written material on two or more separate occasions. It may be quite difficult to keep the different assessments independent in the sense required. Students may repeat aspects of a performance or answers to a series of questions because they remember their behavior on the previous occasions, or it may simply be inappropriate to repeat the same interview or performance assessment. Because of this type of problem, it is often necessary to resort to parallel tasks and to design a study that separates error due to student unreliability from error due to task unreliability. A problem is also encountered with products and written material, particularly where one unique item is being assessed. It may not be possible for the student to "repeat" the performance, in any true sense. Although assessment error due to student unreliability is often difficult to assess, it is also probably less troublesome than judge or task reliability.

5. Separating Sources of Error Statistically. In special cases it may be possible to separate sources of unreliability and identify the magnitude of error in each source. Normally, this would require a very carefully planned experiment in which students, judges, and subtasks are sampled on a systematic basis. A statistical procedure, known as analysis of variance, can then be used to break up and isolate the sources of error in assessment. An expert with a thorough knowledge of experimental design and statistics should be involved in the planning of such an experiment.

Assessing the Validity of Assessment

Evaluating the validity of an assessment procedure also requires the establishment of special experiments or data collection procedures. In one sense, validation can be thought of as a verification process in that the results of assessment are verified with information relating to the competence or knowledge measured. Following are some examples of approaches to validating assessment procedures.

1. Predictive Validity. In cases where credit is assigned for a prerequisite, it may be possible to design an experiment where students would be rated as to their readiness for taking an advanced course by a given assessment procedure. For experimental purposes a sample of students could be allowed to enter the advanced course regardless of the rating they obtained in the assessment. The

assessment ratings could then be correlated with advanced course performance.

2. Concurrent Validity. It would be possible to choose a sample of students ranging considerably in documented learning experiences in a given area. An assessment procedure could then be used to measure the learning or competence of each student without making this evidence available to the judges. The results of the assessments could then be compared with the documented evidence. This method might be appropriate for an oral test or a performance test and would be particularly appropriate for product assessment. It would clearly not be appropriate for an interview, when one of the functions of the interview is obtaining information about learning experiences.

3. Content Validity. An empirical approach to content validity can be taken by having two independent teams of experts devise an assessment procedure given the same set of initial criterion standards. A sample of students could then be assessed with each procedure, and the overall assessments obtained with each procedure could be related. A high relationship would indicate content validity for both procedures.

4. Special Investigations of Accuracy. For some procedures, such as interviews, it may normally be very expensive to check into the accuracy of all the information obtained from a student during an assessment interview. However, it would be possible on an experimental basis to use extensive checking procedures and thereby obtain an estimate of the accuracy of information obtained during the interview.

Judgmental Assessment of Assessment

In order for assessment to be judged without conducting an actual experimental study, it is first necessary that a very careful documentation be made of the entire assessment process. Whether the procedure is a highly structured performance test or an unstructured interview, this documentation means that criterion standards should be written out, a plan for evaluation should be explicitly written, and how the plan for evaluation relates to the criterion standards should be made explicit. For interviews and performance assessment, electronic recording plus some written record of observation should be made. Finally, the assessor should state how the decision or final overall judgment relates to the observations and criterion standards. Given such a systematic record of an assessment procedure, it is possible to have several experts judgmentally assess the internal consistency of the assessment and the relevance of the final judgment. The experts assessing the assessment should include at least one expert in the content area and at least one expert in measurement.

RECOMMENDED BASIC STEPS FOR CONDUCTING ASSESSMENTS

What follows is an attempt to summarize in a general way a set of basic steps for improving assessment applicable at different levels to any assessment situation. In

developing, revising, and monitoring an assessment procedure, it will be necessary to involve several different points of view—that of the expert in a specific substantive area, that of the expert in the methods and problems of assessment, and that of the students who will be assessed.

1. Establish clear criterion standards. This should be done in explicit, observable terms that have a clear meaning for both the assessor and the student.
   a. Begin with general definitions of competence or learning that can be applied to all fields.
   b. Define specific learning outcomes that should be assessed.
   c. Write definitions that can be specifically applied to the area of learning in which the student is seeking credit.
   d. Identify critical behaviors or indices that relate to different levels of competence or learning. This is particularly important when the assessment can lead to varying numbers of credits.
2. Select and structure an assessment procedure that will elicit a sample of observable behaviors, indices, or other evidence representative of the criterion standards.
   a. First determine whether the proposed assessment procedure is content valid. Use a systematic approach (such as that illustrated in Figure 2).
   b. Revise and extend the procedure as necessary to ensure that all important objectives are covered.
3. Plan the administration of assessment.
   a. If more than one assessment task or modular component is used, the order in which these components are administered should be specified.
   b. Make sure that instructions for both students and assessors are clear.
   c. Make sure that provisions are made for the required space, time, and equipment.
   d. Choose assessors carefully and make sure that they understand the procedures fully.
   e. Provide for recording of observations electronically, through rating scales, checklists, or detailed note-taking.
   f. If possible, try out the assessment method with one or more students. The tryout could involve an actual student or could utilize role-playing techniques.
   g. Based on the tryout, revise and make final procedures for administration.
4. Ensure that judgments are based on what has been observed.
   a. Try to have at least two assessors.
   b. Familiarize assessors with common types of errors made in assessment.
   c. Provide assessors with observational aids such as rating scales or checklists.

d. Establish the requirement that all assessors describe in a brief written report the observations they made and how those observations relate to their assessment of the student's competence or learning.

5. Record and report the results of assessment.

   a. Scores should be reported in permanent records, including student records and those designed for data collection, research, and reporting.

   b. Set up procedures for reporting and interpreting test results to students, which would include:

      (1) a written report to the student containing score interpretations.
      (2) an interview covering interpretation, diagnosis of learning problems, and suggestions for further study or practice on performance skills.

6. Monitor assessment procedures.

   a. Detailed records and reports can serve as inputs for evaluating the effectiveness of assessment. Someone with methodological experience and expertise in assessment should periodically monitor such records to determine whether problems exist and, if so, what can be done to correct them.

   b. A second monitoring procedure would involve having a methodological expert directly observe assessments. Direct observation may suggest problems not discernible through written reports.

   c. A more elaborate procedure for monitoring assessment would involve conducting empirical studies of reliability and validity, such as those suggested in the preceding section.

Bibliography On Methodological Problems And Issues

METHODOLOGY AND PRINCIPLES OF MEASUREMENT

Anastasi, A. *Psychological testing.* New York: MacMillan, 1972.

Cronbach, Lee J. *Essentials of psychological testing.* New York: Harper, 1960.

Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

Horrocks, J. E. *Assessment of behavior.* Columbus, Ohio: Charles E. Merrill Books, 1964.

Mehrens, W. A., and Lehmann, I. J. *Measurement and evaluation in education and psychology.* New York: Holt, Rinehart and Winston, 1973.

Nunnaly, J. C. *Psychometric theory.* New York: McGraw-Hill, 1967.

Thorndike, Robert L. *Personnel selection.* New York: Wiley, 1949.

Thorndike, Robert L. *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 1971.

EXPERT JUDGMENT IN ASSESSMENT

Barret, R. S. *Performance rating.* Chicago: Science Research Associates, 1966.

Kavanaugh, M. The content issue in performance appraisal: A review. *Personnel Psychology,* 1971, vol. 24, pp. 653-668.

Knapp, Joan, and Sharon, Amiel *A compendium of assessment techniques.* Princeton, N.J.: CAEL, Educational Testing Service, 1975.

Lopez, Felix M. Measuring human performance. Chapter 8 in Felix M. Lopez, *Evaluating employee performance.* New York: American Management Association, 1968, pp. 163-185.

Menges, Robert J. Assessing readiness for professional practice. *Review of Educational Research,* 1975, vol. 45, no. 2, pp. 173-207.

Smith, P. C., and Kendall, L. M. Retranslation of expectations. *Journal of Applied Psychology,* 1963, vol. 47, pp. 149-155.

Whisler, T. L., and Harper, S. F. *Performance appraisal.* New York: Holt, Rinehart and Winston, 1962.

Willingham, Warren W. Critical issues and basic requirements for assessment, in Morris T. Keeton and Associates, *Experiential learning: Rationale, characteristics, and assessment.* San Francisco: Jossey-Bass, 1976, pp. 224-244.

III. The Interview and Related Procedures

Judith Pendergrass, Jane Porter Stutz, and Richard Reilly

Of the four specific techniques discussed in this *Handbook*, the interview and its variants are perhaps the most commonly used for the assessment of experiential learning.

When an interview or, for that matter, any measurement technique is part of the assessment process, its exact functions depend upon the institution where it is to be used and can be considered only in the context of institutional goals. Thus, the purpose of an assessment interview at one college may be quite different from the purpose of a similar interview at another institution. Similarly, the suggestions offered later in this chapter for conducting an interview or other assessment will be applicable at some institutions and not others, and within a single institution in some situations and not others. It is not the intention of this chapter to dictate what the purpose of an assessment technique should be, but rather to describe some functions that such assessment might serve and to provide some common-sense hints that may prove useful in assessment. It must be left to an individual institution to determine what is appropriate in its particular case.

FUNCTIONS OF THE INTERVIEW

Some functions of interviews that relate to but do not actually constitute assessment are

- obtaining information
- communicating information
- influencing or changing behavior
- diagnosing problems
- verifying documentation

For example, the exploratory interview is often used at institutions that credit experiential learning to gather information about a student's background, interests, goals, etc., from which options can be suggested that the student may follow. In this type of situation the interviewer may be said in some vague sense to be "assessing" the student, but such an interview does not constitute assessment in the sense of measuring a specific learning outcome in terms of what it is worth toward a college degree. The exploratory interview can be considered one type of "counseling interview," and interviews of this sort—in which progress is discussed and options suggested—may be employed throughout the student's college career. Counseling interviews may also be used to prepare a student for a panel interview or other assessment or in some way to influence a student's behavior.

A very common use of the interview is verification of documentation, which is not assessment in the strict sense. If, for example, a student brings in a letter of reference, the interviewer usually asks some questions about the experience refer-

red to in the letter. In this case, there is no product or performance to be assessed directly; there is merely description.

Interviewing for diagnostic purposes is also common in education. One example is the holistic interview conducted to determine whether there are gaps in a student's education. A counseling interview that results in recommendations for next steps is also, in a sense, diagnostic.

In terms of assessment through oral procedures—the focus of this chapter—the interview functions suggested so far are secondary. The primary function of an assessment interview is to obtain information or evidence so that a well-informed judgment can be made regarding a student's competence or knowledge. An oral procedure may be used to assess knowledge or skills directly, as in an oral examination, or to supplement another means of assessment, as in the case in which an interview provides background and context for assessment of a product such as a short story or a musical composition.

Many of the hundreds of research studies that have been done over the years have concluded that interviews, particularly unstructured interviews, lack reliability and validity. It is encouraging to note, therefore, that the CAEL validation[1] study of assessment interviews reported both reliability and validity at acceptable levels. In brief, this study found that there was substantial agreement (reliability) between an interviewer's and an observer's judgments of an interviewee's counseling competence. Moreover, there was a significant relationship (validity) between assessments made with the interview and independent assessments made with a role-playing technique. These results provide perhaps the first systematic research evidence to support the use of interviews to assess experiential learning.

TYPES OF ASSESSMENT INTERVIEWS[2]

An interview can be defined as an oral exchange between two or more people that differs from an ordinary conversation in that there is some structure, there is a goal or purpose, and the roles of the participants are clearly defined. Four more or less distinct types of assessment interviews can be identified: the oral examination, the one-to-one interview, the panel interview, and the leaderless group discussion, within each of which a range of structure is possible.

In each case, a detailed schedule may be drawn up for conducting the assessment, or a list of discussion topics may be used. Many institutions utilizing these techniques employ the whole range, determining the amount of structure necessary on the basis of the knowledge to be measured, the evaluator's wishes, and the skills of the individual student. Consequently, variation in specificity of detail exists among the four types of interview techniques as well as in the ways that institutions design them for the evaluation of a particular competence.

Oral Examination. An oral examination is an interview, usually highly structured, in which questions are planned in advance and relate directly to the competence being assessed. Often preset standards determine the assessment decision re-

---

[1]See Warren W. Willingham and Associates, *The CAEL Validation Report*, Chapter IV.

[2]For another discussion of types of interviews used for assessment, see *A Compendium of Assessment Techniques*, by Joan Knapp and Amiel Sharon.

garding the student's competence. Oral examinations are used in fields like medicine to determine certification and licensing, and in fields like counseling, where interpersonal competence is assessed. The oral examination is a relatively objective demonstration of competence or knowledge. It provides an opportunity for the examiner(s) to ask for clarification or justification, or to judge the manner and content of the reply to a direct question, and it has great potential for use in conjunction with performance testing. The major disadvantage of the oral exam is that it is time-consuming and costly (compared to a group paper-and-pencil exam).

One-to-One Interview. When one interviewer questions one student, the interviewer can devote full attention to a single individual and can, in some cases, adapt the questioning (style, technique, and content) accordingly. There is a range of structure in one-to-one interviews. A very structured interview makes use of a "schedule"—a list of specific questions to be asked in a predetermined order. In a semistructured interview, the interviewer knows exactly what information is required but varies the wording and sequence of questions for maximum effectiveness with individual students. In the least structured one-to-one interview, the interviewer determines in advance the areas or topics to be covered but varies the questions with the specific situation. Since interviews are dependent to some extent upon personality, they are less reliable than other forms of assessment; obviously structured interviews are more reliable than unstructured interviews. An advantage of the one-to-one interview is that it may require relatively little preparation compared with other procedures. The disadvantages are, again, the cost and the time required. Many experiential learning situations, ranging from foreign travel to volunteer community work, lend themselves to assessment by one-to-one interviews.

Panel Interview. In a panel interview, several interviewers question a single student during one session. This is, of course, less time-consuming for the interviewee than being interviewed by each examiner separately, and the interviewers can benefit collectively from each other's questions and group discussion. The panel interview can also serve as a simulation of a situation where the individual must account to a group. On the other hand, the panel interview is subject to "contamination"; i.e., interviewers' comments and questions may influence each other's views and lead them to a different evaluation from one that they would have arrived at individually. The panel members also need to take care that the interview does not become a "trial by jury." Finally, the panel interview may be especially anxiety-producing for the student; although this may be an advantage if the competence to be assessed involves ability to act or think under stress, normally students can be assessed best if they are at ease.

Leaderless Group Discussion. In this type of oral assessment several examinees are asked as a group to carry on a discussion while the examiner(s) observe. No leader is appointed, and the examiners do not enter the discussion. The leaderless group discussion (LGD) is often used to assess competences related to leadership skills. As with other oral techniques, there is a range of structure possible within LDGs. At one extreme, a very general topic may be presented for discussion; at the other extreme, each participant may be given

detailed background information and asked to play a specific role. The advantages of the LGD are that it can elicit a type of interpersonal behavior with peers that is not readily observable in other formats, and it can provide a direct learning experience and feedback for the student. The disadvantages are that the amount of structure that can be imposed is limited, and as a result the validity and reliability are also limited. Furthermore, several assessors are needed.

DEVELOPING THE INTERVIEW

Key Steps

This section offers specific recommendations for the development of the interview itself.[3] In each case, the recommendations are fairly elaborate and assume extensive resources in terms of personnel and time. It is recognized, however, that resources are often quite limited and that it may not always be desirable or feasible to carry out all of the steps. Each of the suggested steps will usually have to be considered in light of the assessment problem and the existing resources. Carrying out all of the steps would be especially desirable for assessment situations that are likely to be repeated many times.

1. Establish Criterion Standards and Determine Appropriate Levels. The first two steps in the development of an assessment interview are common to each of the specific procedures and essentially involve the establishing of criterion standards discussed in Chapter II. That is, the experts should first determine the levels of competence in the specific area to be assessed by writing fairly objective definitions (as in Figure 1) and then, as a second step, should define those levels in objective observable terms. These two steps will then guide the development of the complete interview.

2. Develop the Content. Once criterion standards have been established and explicitly defined, the interview schedule itself should be drawn up. Questions should be designed so that they are directly related to the criterion standards. A representative sampling of questions should be developed to cover the subject matter. Questions may be of varying difficulty and should be arranged systematically, such as moving from the easier to the more difficult ones.

3. Develop Observational and Recording Procedures. A procedure for recording and rating students' responses must be developed in advance of the administration of the procedure. A rating scale or checklist, with space provided for comments, may prove to be most useful. The use of such a recording instrument is especially important in a leaderless group discussion. Behaviorally-anchored rating scales, such as the one presented in Figure 4, could be developed or checklists of behaviors could be prepared. In each case, the rating scale should be closely tied to the criterion standards.

---

[3]For a detailed discussion of developing an interview for assessing interpersonal skills, see Paul Breen, Thomas F. Donlon, and Urban Whitaker, *Teaching and Assessing Interpersonal Competence–A CAEL Handbook*, Chapter VII.

4. Tryout, Evaluation, and Revision. All questions, materials, and rating scales should be tried out with a small sample of students and more than one judge. This will allow ambiguities to be cleared up and unnecessary questions dropped. Of course, as has been mentioned, an elaborate try-out procedure may be impossible. In general, however, some attempt should be made to determine if the assessment process elicits responses that are adequate as well as relevant to the criterion standards.

5. Development of Final Products. A final set of questions or discussion topics, guidelines for administration, and a recording form should be produced. In the case of a leaderless group discussion, a set of background materials for students should be included.

Designing The Interview Schedule

Establishing Content and Criterion Standards. First, the persons responsible for designing the schedule must determine the categories of knowledge and skill fundamental to the area of learning to be assessed. In some institutions several consultants in the field may aid in this process, and the responsibility for completing the interview schedule itself may be given to a member of a special assessment staff or to another faculty member. This approach is particularly common in cases in which rather structured schedules are being developed or several evaluators will be using a particular schedule over a period of time.

For example, in developing the interview for measuring leadership skill for a volunteer organization (included at the end of this chapter), consultants identified twelve such categories. Although these categories differ somewhat in scope, each is actually composed of several facets, and it is important that the questions used within the interview be designed to cover some but not necessarily all of the subcomponents in each of the categories. Although there are not required answers that are necessary to demonstrate competence in the area, the student should possess two abilities that cut across all twelve subjects. These are the ability to analyze behaviors and the reasons for them and the ability to compare organizations and activities in order to determine the generalizability of the knowledge gained. It is important to note at this point that such an interview is designed not to validate the student's participation in a set of experiences, but rather to assess the outcomes of those experiences.

Following agreement on the basic categories of knowledge and skills, a specific set of criterion standards is derived that will later serve as the basis of a rating scale to be used during the administration of the interview itself. For example, consultants derived the following standards for leadership competence in volunteer organizations. A student should:

1. Understand and apply the goal-setting process; understand the organization's goals and evaluate them in setting priorities.
2. Understand and apply the processes of organizing new organizations and maintaining established ones; evaluate similarities and differences between the two types of processes.
3. Understand the structure of the organization and apply this knowledge by

analyzing and maintaining the relationship between the structure and the organization's goals and central identity.

4. Understand the use of a constitution and its effects on the directions taken by the organization; understand most provisions of this organization's constitution, if any.
5. Understand own leadership style and skills and apply this knowledge in a variety of situations within the organization, and evaluate outcomes.
6. Understand the organization's sources of revenue and fiscal management processes; apply this knowledge in determining and evaluating alternate sources of revenue and accompanying management techniques.
7. Understand the role of various types of communications in performing leadership functions and apply this knowledge in analyzing their variations in different organizations; evaluate own use of communications skills in relation to a specific organization.
8. Understand, apply, and evaluate the value and techniques of using public communications media in achieving an organization's goals.
9. Understand a variety of sources of information and community resources and apply this knowledge in locating and evaluating them in order to complete a task.
10. Understand sources of decision making external to the organization and apply this understanding by analyzing the relationship established between them and the organization.
11. Understand the civic arenas in which volunteer organizations may operate and apply this knowledge by analyzing the arena in which this organization operated and one's own role in being responsive to that arena.
12. Understand and apply the techniques of supervising volunteer and paid staffs.

Determining Appropriateness of Interview Technique. The next step is to decide whether the interview is an appropriate assessment technique for measuring a particular body of knowledge. This determination should be made in light of the criterion standards that have been established. An interview alone may be adequate to assess the competence, though in some cases an additional method, such as a product assessment, should be used.

The interview is especially useful in measuring knowledge of leadership techniques because problem-solving situations may be posed to determine the generalizability of the knowledge. The interview also allows the assessor to focus the interview on special areas of competence as they surface during the course of the interview. Frequently the specific direction cannot be fully anticipated before the interviewer and student become acquainted and the interviewer becomes more fully informed of the student's experiences.

Furthermore, the affective dimension of the competence area is a significant one to measure since it frequently plays a substantial role in determining the direction that a leader of an organization chooses to take and the result of his or her endeavors. The interview is particularly effective in evaluating a student's values and attitudes.

Developing the Question Series. Next a set of questions—the interview schedule—is drawn up for use in measuring the knowledge and skill being claimed by the student. Using a somewhat structured schedule aids in achieving consistency among the interviewers who will be assessing the competence. It also promotes more thorough coverage of the subject as well as similar treatment of all students claiming this knowledge. In many cases time and cost prohibit the development of a complete schedule.

In developing the schedule it is essential to work within the confines dictated by the criterion standards. Each category should also be covered thoroughly, although it is not necessary to separate questions by categories of knowledge. It may sometimes be preferable to design questions that touch on several components at once. This has the advantage of allowing the student to demonstrate understanding of the relationships among various facets of the competence.

Following the development of the schedule, a rating scale should be designed that is based on the criterion standards as well as the schedule itself. An example is included at the end of this chapter. The scale should be easy to complete and useful in making final credit recommendations. Space should be allowed for comments.

Key Questions for Use in Interviewing. The following types of questions may prove particularly appropriate for measuring experiential learning. In general these questions emphasize the importance of the student's articulation of the knowledge she or he has gained. Though similar questions could well be used for evaluating more classic academic areas, adjustments should be made in order to solicit fewer personal perspectives and more specific information. In this case the range of acceptable responses would not be as great. Questions may require the student to:

1. Recall activities and evaluate their priority and value.
2. Analyze issues in a given dimension and evaluate his or her ability to deal with them.
3. Draw comparisons among experiences.
4. Use hypothetical situations to demonstrate problem-solving skill within a given context.
5. Identify trends or systematic changes in his or her actions as well as the reason for them.
6. Analyze his or her own strengths and weaknesses with regard to specific situations or sets of behavior.
7. Analyze his or her basic operating style.
8. Identify theories that are consciously being applied.

Reviewing for Content Validity. Finally, the interview schedule should be examined for content validity, ideally by the consultants who first served in its formation as well as by other experts in the area. This allows the institution to check to be sure that the judgments of the initial subject-matter experts have been maintained. It is also beneficial to test the instrument with persons who have fresh perspectives. In many cases the interviewer will be the sole subject-matter expert

involved in designing and administering the schedule, but it is still most desirable to have another person experienced in the field review the schedule before it is put into frequent use.

CONDUCTING THE INTERVIEW

Preparation

The first step in preparing for an assessment interview is the identification of the assessor. Ideally, the interviewer should have expert knowledge of the field as well as interviewing experience. Some institutions use a pair or a team of interviewers in which both subject-matter expertise and interviewing skills are represented. Of course, the "ideal" situation is not usually at hand, but the person who conducts the interview should in any case be cooperative, flexible, tolerant of different kinds of unique experiences, and willing to use guidelines. (For suggestions on identifying people with subject-matter expertise, see Chapter I.)

Prior to the interview itself, it is desirable for the interviewer and the assessee to meet each other, or if that is not possible, to talk on the telephone. The basic purpose of this conversation is to familiarize the assessor with the student's situation and to give the student an overview of what can be expected during the assessment. In addition, criterion standards can be agreed upon at this time, and the areas to be covered can be explained so that the student can reflect on topics he or she wishes to include in responses. This approach often produces more thorough answers while reserving the elements of fresh investigation and analysis of knowledge for the interview itself. Some interviewers also choose to discuss the rating scale. This meeting before the interview gives the assessor and the student a chance to become acquainted, if only briefly, and can serve to reduce the student's anxiety at the time of the assessment. At this time the details of time, place, and length of the interview can also be finalized.

Preinterview preparation for the assessor involves planning the interview and adjusting the interview schedule to the particular student's situation. For example, in the case of the leadership competence schedule discussed earlier, the principal organization(s) in which the student has participated should be reflected throughout the interview. This may necessitate the addition of some specific terms or information to some of the questions. This prototype was designed with the understanding that the student would be asked to draw conclusions most heavily from the principal organization in which he or she had taken a leadership role, but that incorporation of experiences from other organizations, if any were applicable, would be encouraged. The student would also be urged to hypothesize regarding other situations based on knowledge of the one organization in which he or she had participated extensively.

In administering a prototype interview, it may be necessary to eliminate some items that are not appropriate to the experiences of the student. Obviously, there is a wide variety among students' learning experiences and the specific skills derived from those experiences. In order to be most useful as an assessment technique, the interview schedule must then reflect the peculiar set of knowledge and skills that the student brings to the interview. However, the core criteria, established by the institution to be indicative of learning or competence within this area, should

remain constant among various students. Consequently, the basic interview schedule and components of the competence area should also remain relatively constant.

Some Practical Suggestions

A number of handbooks and manuals have been written on how to interview, particularly in the area of employment interviewing. Several of them are included in the bibliography at the end of this chapter. This section presents those "helpful hints" that apply particularly to assessment interviews. There is nothing esoteric or even technical about these suggestions; most of them are merely common-sense ideas, but awareness of them can improve the effectiveness of the interview.

Interviews or other oral assessment procedures of any sort are best conducted in privacy with as few interruptions and distractions as possible. It is a good idea to find a room where voices cannot be overheard and to arrange to have no phone calls or visitors during the course of the assessment. The interviewer should be sure to arrive on time and come prepared with copies of both the interview schedule and the rating scale so that the grid can be filled out as soon as the interview is completed. If a panel is involved, copies should be provided for each member. When feasible, it can be helpful to conduct practice interviews and to complete practice rating forms before conducting a real interview.

The first few minutes of an interview or oral examination are important because they set the tone for all that follows. The student will be able to demonstrate competence best if he or she is at ease, and the interviewer can help by creating as relaxed and pleasant an atmosphere as possible. This can be accomplished by greeting the student pleasantly and giving the distinct impression that it will be easy to talk with one another. Some manuals suggest a few minutes of "small talk" to help reduce the student's anxiety. The interviewer should have the flexibility necessary to adjust the schedule to his or her own style as well as to the student's. Such adjustment is necessary for maintaining a smooth flow within the interviewing process as well as a comfortable situation for the student. This, in turn, is essential in order to maximize the effectiveness and efficiency of the assessment technique. It is also important to clarify during the first few minutes how long the interview will last, what will be covered, and what is expected. Although this ground may have been covered in a preinterview conversation as mentioned earlier, it is a good idea to review this information at the start of the actual interview.

An important aspect of the interview or oral exam, and one which is often ignored, is how to ask questions so that the student has maximum opportunity to demonstrate knowledge/competence without cues from the interviewer. The following suggestions may prove useful in this regard:

- Choose positive wording.
- Beginning with the first question, establish the pattern of the student's doing most of the talking.
- Make use of the calculated pause; if you don't say anything, the student will often elaborate.
- When possible, use a comment instead of a question. Try to keep the interview from looking like a cross-examination.

- Phrase questions carefully so that they are clearly understandable, so that you don't suggest the answer you're looking for, and so that you don't cause the student to lose face.
- Always have a clear purpose in mind when you ask a question; i.e., don't ask unnecessary questions.
- Use examples to help elicit more complete responses. Many students need that additional assistance in developing a framework or context within which to place their answers. Such examples may be included in the original outline or added for the particular situation.

Although a time limit is usually set for the interview as a whole, it is not necessary to spend a specific amount of time on each section. Such flexibility allows the interviewer freedom to direct the questions in such a way as to elicit the greatest amount of information possible in those areas of greatest competence and to choose those questions in each category that are especially appropriate to the particular student's own experiences. Consequently, some questions included in the original schedule may not be used during the actual administration. Adjustments may also be made in specific questions to make them appropriate to the student's level of sophistication in each category of knowledge.

Questions can be used to remind the student of omitted parts of earlier responses, to get further information or probe more deeply, or to clarify the meaning of earlier remarks. It is the responsibility of the interviewer to help the student to be as definitive and specific as possible and to clarify ambiguous responses. Checking on answers that are unclear to the interviewer is important, but it should be done without the use of leading questions that reveal the desired reply.

Another important element in the interview situation is control. The interviewer needs to maintain control of the interview in order to utilize time efficiently and economically and to insure proper balance and adequate coverage of each area.

Control can be maintained by:

- Being systematic—for example, by following an interview schedule carefully.
- Exhausting one area before going on to the next (if you find you have forgotten something, do not interrupt and go back in the middle or another area).
- Pacing the interview, allowing enough time, not dawdling.
- Avoid awkward pauses, although "calculated pauses" can be used judiciously.
- Focusing attention on the issues in question.
- Pushing the student along with a question if he or she goes into too much irrelevant detail or gets off the track. Anticipate the point where you can interrupt, and do it with what Fear[4] calls "lubrication," i.e., positive comments.

The interviewer can further help the student by being responsive, for example, by really listening; by *showing* that he or she is really listening (by nodding, saying

---

[4]Richard A. Fear, *The Evaluation Interview*, p. 74.

"I see," etc.); by making use of facial expressions, intonation, gestures, postures; and by giving frequent "pats on the back" or supportive comments. Although it may be necessary to tell the student that an answer is wrong, one should avoid letting the interview degenerate into an argument. This supportiveness generally results in a more comprehensive response from the student and consequently a more thorough assessment.

There are two schools of thought as to whether or not notes should be taken during an interview. The argument against note-taking is that it increases student anxiety and can lead to loss of rapport. In the case of an assessment interview, however, it seems absolutely essential that notes be taken for the sake of accuracy. The risk of increasing anxiety or losing rapport may be lessened by telling the student in advance that notes will be taken. Tape-recording the interview is sometimes substituted for or used in addition to note-taking and has the advantage of enabling an additional assessor to evaluate the outcomes of the interview after it is completed. It is important to notify the student before beginning the interview if a tape recording will be made.

Pitfalls to Avoid

Because interviews and other oral assessment procedures are inevitably affected by the personalities of the people involved, they are particularly vulnerable to the problems associated with validity and reliability. Although there is no way to control completely the subtle effects of interpersonal interaction, the interviewer can make more objective decisions by keeping in mind the sources of error in assessment discussed in Chapter II as well as the following pitfalls.

The greatest danger in oral assessment is that some sort of interviewer bias will creep in and affect the outcome. In addition to the halo effect discussed earlier, stereotypes and interviewer expectations are another source of bias. The interviewer must be careful to avoid jumping to conclusions and prejudging a student (either favorably or unfavorably) on the basis of any characteristic. Here are some examples:

- *Age.* The interviewer may be reluctant to give credit to a student who is only 18 or predisposed to give credit to one who is 40. It is easy, but wrong, to assume that since the latter is more than 20 years older, he or she automatically knows more about a given subject or has reached a higher level of competence at whatever is being assessed.
- *Sex.* Traditional male and female roles should not stand in the way of fair assessment; for example, a man may legitimately seek credit for experience as a nurse.
- *Race.*
- *Past experience.* If a middle-aged student claims 10 years of experience as a volunteer social worker, his or her competence still needs to be assessed before credit is awarded. In other words, credit is to be awarded for knowledge, not for experience per se.
- *Dialect.*
- *Voice quality.*

- *Pronunciation* (or *mis*pronunciation).
- *Posture*.
- *Stance* (for example, aggressive or meek).
- *Appearance*. If dressing neatly is not related to the knowledge being assessed, then a sloppily-dressed student should not be penalized for his or her appearance.

In addition, the interviewer should beware of prejudice about the *area in which the student is requesting credit*. The male nurse example cited above could fit this category as well. Finally, the interviewer should be sure that the *implicit long-range use of the assessment* does not affect the interview outcome. *Assumptions* about the use to which the student will ultimately put the assessment may be false. If, in the context of the individual institution, different standards are appropriate depending on the long-range use of the assessment, then the assumptions should be checked and the long-range goals clarified. It may in some contexts be appropriate, for example, to have different artistic standards for the student who aspires to be a professional painter and for the student who plans to work in advertising.

In addition to interviewer bias, another pitfall to avoid in oral assessment is misunderstanding. If the interviewer feels at any time that there may be a misunderstanding, on either side, he or she should check and clarify. Misunderstanding on the part of the student can be minimized by careful and clear wording of questions. A final caution—the interviewer must avoid letting the student take over the interview.

ANALYZING THE OUTCOME OF THE INTERVIEW

After conducting the interview, the evaluator or panel should review the student's responses to individual questions or to sections of the schedule. In general, it is helpful to arrive at a summary decision regarding the level of knowledge demonstrated and the overall quality of the student's discussions.

The level of student competence will, in most cases, vary from item to item and from area to area. One of the most difficult tasks of the assessor is to balance these variations and arrive at a summary judgment. It is helpful to support overall evaluations with specific, representative examples of stronger and weaker responses. These specific examples will also be useful in discussing the results of the interview with the student.

Some examples from an actual assessment interview may help to illustrate this point. The examples were taken from a structured interview developed to assess leadership skills. The interview schedule and rating scales used were developed around 12 dimensions of leadership and are appended to the end of this chapter. How the assessor can back up a summary evaluation is shown in the three examples which follow.

1. The student was able to see the relationship among the components of this competence area. Consequently, responses combined information regarding several areas. The interviewers had the flexibility, then, to eliminate questions that had been answered in another context and to encourage the student to draw comparisons among responses when that was appropriate.

2. It was apparent that her levels of knowledge and skill varied among the 12 dimensions. She was clearly at the evaluative level in the use of leadership skills but had only basic knowledge of external power structures. She will need to complete additional study in this area in the future.
3. One of the most outstanding qualities of this student's responses was her ability to analyze, in ways new to her, her own past behaviors and evaluate their outcomes. She was then able to make recommendations regarding procedures that might have been used differently or operating styles of her own that she now feels might have been made more successful.

Following a review of the student's responses, the rating scale should be completed, and comments, either for the student or the advisor should be prepared. Whenever possible a copy of the completed rating scale should be given to the student. Next, a judgment must be made regarding the amount of credit or recognition the student will receive, although some items in a schedule will allow for various interpretations by judges. However, tryout activities can help predict where variations may arise and permit some preparation of evaluators.

Institutions may adopt different approaches or criteria for awarding credit or formal recognition of competence. A comprehensive discussion of guidelines for use in awarding credit is included in *Assessing Prior Learning–A CAEL Handbook* by Joan Knapp. Each of these guidelines is applicable to the analysis of the results of an evaluation interview.

Following the completion of the interview process, the judge or evaluation panel should discuss the evaluations with the student. During this time the specific outcome—the level of formal recognition or credit awarded—should be made clear. Then, too, it is beneficial to the student if some of the responses themselves are discussed and areas for improvement or further study are identified. It is important that this discussion center on the criterion standards previously developed and that it be clear that those standards were the bases upon which the evaluation was made.

PROTOTYPE INTERVIEW SCHEDULE

The following structured interview schedule and rating scale were developed at Metropolitan State University. They were designed with the help of specialized consultants and reflect the context of this particular institution and the orientation of these consultants.[5]

Educational institutions that recognize knowledge gained through experiential learning are often asked to assess competence in leadership of volunteer organizations. Consequently, this area was selected for the development of a model. However, similar procedures could be used in designing interview schedules to measure a diversity of experiential learning, including more traditional academic subjects.

---

[5]Metropolitan State University expresses special thanks to: Sandra Hale, Elmer John, and Richard Leider.

Prototype Interview for Assessing Competence in Leadership of a Volunteer Organization

I. Goal Determination

1. What long-range and short-term goals form the basis of the principal organization which you have served in a leadership capacity?

   To what extent were you a determining factor in establishing or redirecting those goals?

2. In light of these long-range and short-term goals, what action priorities have been established?

   What has been your role in establishing these priorities?

   What factors do you see as the primary determinants of appropriate priorities for directing the work of a volunteer organization?

3. Is there ever a reason for disbanding or substantially altering the objectives and functions of an organization such as this?

   If so, under what circumstances?

   To what extent have you been involved in such a procedure, and how would you evaluate the success of the changes?

II. Organizational Processes

1. What do you see as the major problems to be considered in organizing a new or existing group operating on a volunteer basis?
2. What steps would be particularly essential in developing such a new organization?
3. If you have undertaken the organization of a new group, describe the activities in which you engaged and the decisions which you made during that process. For example, in what way did you determine and establish the central identity, purpose, and functions of your group?
4. In what ways do organizational procedures and general maintenance of the organizational identity differ between a firmly established organization and one which is in its foundation stages?
5. In what ways do the various objectives and functions of a group cause divergent concerns on the part of the membership of the organization to arise?

   In what ways have you dealt with these types of concerns in interpreting, or adjusting, the identity or central focus of the group's activities?

III. Organizational Structure

1. Describe the basic structure of your organization, including the various levels of leadership, their responsibilities, and any subcommittees, along with their functions and the rationale for the formation of those committees.
2. Explain the relationship between this structure and the organization's long-range goals.

   To what extent would or have changing priorities caused adjustments in the organization's structure?

3. In what ways does the size of the organization influence the development and maintenance of a specific organizational structure—for example, in what ways might responsibilities be divided between a governing board and the functioning staff of an organization?

4. How does this structure affect the actual operations of the organization? Cite an example from your experience to support your point.

IV. Constitutional Provisions

1. Is your organization based upon a constitution or charter?

   If not, what might have been the effect of such a document on the identity or activities of your organization?
2. How was the constitution or charter of your organization, if any, first designed?

   Have there been substantial changes since the formation of the organization, and to what do you attribute these changes?

   If applicable, evaluate your role in the writing of the constitution or charter and/or the by-laws by which the organization is governed.
3. What issues seem to you to be most significant in facilitating an organization's following its constitution and by-laws?

   Describe any situations, from your own experience, which necessitated interpretation of these written guidelines or problems arising from their use by the membership of the organization.

V. Leadership Patterns

1. Describe what you consider to be your leadership style in directing the organization under discussion.

   What do you see as your most valuable leadership skill in maintaining the motivational level of the membership?
2. To what extent does your mode of leadership use general interpersonal skills?

   Cite some specific incidents, if any, in which your use of such skills was particularly essential and evaluate the results.
3. In what ways do leadership functions and styles need to be adjusted according to varying sizes and functions of organizations?

   Cite several specific incidents during your period of leadership of one organization which particularly tested your leadership model and/or skills.

   How did you resolve these difficulties, and what were the outcomes of any actions taken?
4. Assume that one member of the group attempted to dominate all the meetings by directing the discussion and monopolizing the conversation to the point that members were beginning to cease attending meetings. In what ways would you deal with this situation, keeping in mind the objectives of the organization and the fact that it is a volunteer organization?

   What tools of discussion might be used in such a situation?

VI. Fiscal Management

1. In the leadership structure of your organization, who has been given the responsibility for the financial management of the organization?

   What basic procedures were used?
2. What are the chief sources of revenue for this organization?

   Explain the ways in which that revenue is derived and your role in that process.
3. If this organization were to face a financial difficulty, how would you go about gaining additional sources of money?

   What conditions within the community would have to be considered?

How might various sources of revenue affect functions of the organization or the central identity of its membership?

4. Are there any legal restrictions of which you are aware which would restrict the sources of revenue open to you or the fundraising activities in which you might engage?
5. Has your organization engaged in any types of fund-raising activities?

   If so, describe briefly the chief techniques used. What do you see to be the most crucial elements in completing such activities successfully?

   Why did you elect to use the activities which you have selected?

VII. Communications

1. What other types of publications were prepared and used by your organization?

   What was your role in their preparation?

   How would you evaluate their effectiveness?
2. Which types of writing did you use while holding the leadership role in the organization?

   How might the types of writing or styles used vary with the size, type, or purpose of the organization? (For example, might the target groups for written fund-raising material vary from group to group?)
3. In what types of public-speaking situations were you involved while performing your leadership functions?

   To what extent were other members of the group involved in similar types of speaking situations and why?
4. What other forms of oral communication have you used while fulfilling your leadership role?

   What role did they play in your leadership style?

   How effective do you feel your use of oral communication skills was?

VIII. Use of the Media

1. To what extent is it feasible for a basically volunteer organization to use one or more of the public communications media in enhancing its work?
2. In what ways did you use various public communications media in connection with achieving the goals of your organization?

   Evaluate the outcomes of your own uses of the media.

IX. Uses of Resources

1. What types of information was it necessary for you to obtain in your leadership capacity?

   How did you go about obtaining that information?

   Which sources were most helpful?
2. Assume that you have been asked to seek an external source of funding, such as a grant from a private organization, for the purpose of extending some special service to the community. How would you go about locating and evaluating various sources of funding for your project?

X. External Power Structures

1. With what sources of decision-making influence external to the organization have you dealt?

Describe the levels of responsibility of some groups with which you have had to deal or which you have attempted to influence in gaining decisions favorable to your organization.

Within that structure, where did pressure groups operate, and in what ways have you dealt with those pressure groups?

What were the outcomes of your endeavors?

2. In what ways might the types of influence which could be brought to bear upon decision-making organizations, such as the state or federal government or another community organization, change with the size or overall goals of any given organization?
3. What types of activities have you used to maintain continuing relationships with such sources of influence?

   How does this type of maintenance differ from more short-term but incisive activities undertaken to influence these power bases?

XI. Operating Arena

1. Assume that you have decided to open a new branch of your organization or to expand its membership into a previously untapped area. What steps would you follow in determining what target groups might be most appropriate for this activity and what needs your organization might meet in potential members?
2. Describe the specific civic arena in which your own organization has operated.

   What types of issues, concerns, and needs within the community are you aware of which have influenced the development and direction of your organization?
3. Describe several community arenas in which volunteer organizations might operate and explain the leadership role which you would probably play in light of the objectives of that community group, the limitations of action available, etc.

XII. Staff Supervision

1. To what extent have you had secretarial or clerical help within your organization?

   Has this been paid or volunteer support?

   What has been your supervisory style in maintaining a continuing relationship with clerical staff or other paid workers within the organization?
2. What problems are unique to supervising volunteer workers in an organization?

   What approaches have you used to maintain continuing interest and motivation in completing necessary tasks?
3. What differences do you see in leading an organization which is supported substantially by volunteer work and one in which much of the day-to-day work is done by a paid staff?

Rating Scale for Use in Assessing Competence in Leadership of a Volunteer Organization

Student	Competence Area
Assessor	Date

Instructions:

1. Read the title and all three descriptions before completing the assessments.
2. For each of the items listed below, place a mark on the line at that point which best represents your assessment of the student's demonstration of that knowledge or skill. The vertical lines represent approximate dividing lines between 3 levels of performance on each item.
3. If an item is not definitely applicable to this student, you may omit it. However, please specify below that item the reason for its inapplicability.
4. Indicate specific strengths and weaknesses in regard to each item in the sections reserved for comments.

I. Goal Determination

Understands the organization's chief goals	Understands the organization's goals and can help set priorities	Understands and applies goal-setting process; understands the organization's goals and can set priorities

Comments:

II. Organizational Processes

Understands how this organization was formed	Understands general organizational processes	Understands and applies processes of organizing new organizations and monitoring established ones

Comments:

III. Organizational Structure

Understands some structural features of the organization	Has comprehensive understanding of the structure of the organization	Understands and applies the structure of the organization and its relationships

Comments:

IV. Constitutional Provisions

Understands some characteristics of a constitution	Understands some provisions of this organization's constitution, if any	Understands the use of a constitution and its effect on the directions taken by an organization

Comments:

V. Leadership Patterns

Understands most facets of own leadership style	Understands elements of developing a leadership style	Understands, applies, and evaluates own leadership style

Comments:

VI. Fiscal Management

Understands some of organization's sources of revenue	Understands organization's primary sources of revenue and some management processes	Understands organization's sources of revenue and fiscal management processes; applies this knowledge in determining and evaluating alternate sources of revenue and management processes

Comments:

VII. Communications

Understands general role of communications in leadership	Understands general role of communications and identifies some of own skills	Understands, applies, and evaluates various types of communications in leadership

Comments:

VIII. Use of Media

Understands some uses of the media	Understands and applies some techniques of using the media	Understands, applies, and evaluates the values and techniques of using media in achieving the organization's goals

Comments:

IX. Use of Resources

Understands some major sources of information within the community	Understands the value and techniques of using several major sources of information	Understands a variety of sources of information; applies this knowledge in locating and evaluating community resources to complete a task

Comments:

X. External Power Structures

Understands the existence of external power structures	Understands organization of some external power sources	Understands structure and role of external sources and applies this understanding in relation to own organization

Comments:

XI. Operating Arena

Understands general arena of own organization	Understands several possible arenas for such organizations	Understands the civic arenas in which volunteer organizations operate; applies this knowledge by analyzing arena in which own organization operates

Comments:

XII. Staff Supervision

Understands some techniques of supervision	Understands some differences in supervising volunteer and paid staff	Understands and applies techniques of supervising volunteer and support staff

Comments:

Bibliography On The Interview And Related Procedures

Bass, Bernard M. The leaderless group discussion technique. *Personnel Psychology,* 1950, vol. 3, pp. 17-32.

Bass, Bernard M. Situational tests: I. Individual interviews compared with leaderless group discussions. *Educational and Psychological Measurement,* 1951, vol. 11, no. 1, pp. 67-75.

Bass, Bernard, M. The leaderless group discussion as a leadership evaluation instrument. *Personnel Psychology,* 1954, vol. 7, pp. 470-477.

Berger, Bernard, and Hoberman, Solomon. Notes on the use of group oral tests. *Public Personnel Review,* July 1955, vol. 16, no. 3, pp. 143-147.

Bingham, Walter Van Dyke; Moore, Bruce Victor; and Gustad, John W. *How to interview.* (4th ed.) New York: Harper and Brothers, 1959, 277 pp.

Breen, Paul; Donlon, Thomas F.; and Whitaker, Urban. *Teaching and assessing interpersonal competence–A CAEL handbook.* Princeton, N.J.: CAEL, Educational Testing Service, 1977.

Fear, Richard A. *The evaluation interview.* (2nd ed.) New York: McGraw-Hill, 1973, 320 pp.

Fenlason, Anne; Beals, G.; and Abrahamson, A. *Essentials in interviewing.* (rev. ed.) New York: Harper and Row, 1962.

Fields, Harold. An analysis of the use of the group oral interview. *Personnel,* 1951, vol. 27, no. 6, pp. 480-486.

Gordon, Raymond. *Interviewing: Strategy, techniques, and tactics.* Homewood, Ill.: Dorsey Press, 1969.

Harral, Stewart. *Keys to successful interviewing.* Norman: University of Oklahoma Press, 1954, 223 pp.

Hubbard, John P. The oral examination, in John P. Hubbard, *Measuring medical education: The tests and test procedures of the National Board of Medical Examiners.* Philadelphia: Lea and Febiger, 1971, pp. 93-99.

Kahn, Robert L., and Cannell, C. F. *The dynamics of interviewing.* New York: Wiley, 1957.

Knapp, Joan. *Assessing prior learning–A CAEL handbook.* Princeton, N.J.: CAEL, Educational Testing Service, 1977.

Knapp, Joan, and Sharon, Amiel. *A compendium of assessment techniques.* Princeton, N.J.: CAEL, Educational Testing Service, 1975, 50 pp.

Maier, Norman R. F. *The appraisal interview: Objectives, methods, and skills.* New York: John Wiley, 1958.

Mayfield, Eugene C. The selection interview: A re-evaluation of published research. *Personnel Psychology,* 1964, vol. 17, no. 3, pp. 239-260.

McGuire, C. H. The oral examination as a measure of professional competence. *Journal of Medical Education,* 1966, vol. 41, pp. 267-274.

Morgan, Henry H., and Cogger, John W. *The interviewer's manual.* New York: Psychological Corporation, 42 pp.

Payne, Stanley L. *The art of asking questions.* Princeton, N.J.: Princeton University Press, 1951, 249 pp.

Prien, Erich P., and Lee, Robert J. Peer ratings and leaderless group discussions for evaluation of classroom performance. *Psychological Reports,* 1965, vol. 16, no. 1, pp. 59-64.

Richardson, Stephen A., Dohrenwend, Barbara S., and Klein, David. *Interviewing: Its forms and functions.* New York: Basic Books, 1965, 380 pp.

Sharma, Prakash C. *Interview and techniques of interviewing: A selected research bibliography (1930-1966).* Exchange Bibliography 715. Monticello, Ill.: Council of Planning Librarians, 1974, 29 pp.

Shouksmith, George. *Assessment through interviewing.* Elmsford, N.Y.: Pergamon Publishing Company, 1968.

Sidney, Elizabeth and Brown, Margaret. *The skills of interviewing.* London: Tavistock Publishing Co., 1961.

Wagner, Ralph. The employment interview: A critical summary. *Personnel Psychology,* Spring 1949, vol. 2, pp. 17-46.

Webster, E. C. *Decision making in the employment interview.* Montreal: Applied Psychology Centre, McGill University, 1964, 124 pp.

Whisler, Thomas L., and Harper, Shirley F. *Performance appraisal: Research and practice.* New York: Holt, Rinehart and Winston, 1962, 593 pp.

Willingham, Warren W., and Associates. *The CAEL validation report.* Princeton, N.J.: CAEL, Educational Testing Service, 1976.

Wright, Orman R. Summary of research on the selection interview since 1964. *Personnel Psychology,* 1969, vol. 22, pp. 391-413.

IV. Product Assessment

Ruth Churchill

In many cases the most direct way to evaluate the learning that is either implicit or explicit in a given experience is by means of an appropriate product or products. This is clearly the situation in the creative arts. Students claiming skills in the visual arts, for example, should have produced paintings, sculpture, photographs, films, or other art forms that give evidence of their skills. Writing and musical composition are other examples of arts in which the product itself is important as evidence of learning. In other areas the importance of the product may be less clear-cut, but in many fields products are evidences of the learning of skills and understandings. In science, exhibits and demonstrations are regularly used in science fairs as bases for making judgments about the learning of such objectives as scientific methods of thinking, the ability to present scientific data clearly and to apply scientific ideas to new situations, and the technical skills in constructing equipment effectively and safely. In teaching, a curriculum unit developed to teach certain knowledge, understandings, or skills by a particular method to a given group of learners is a product that reflects an understanding of educational philsophy, a sensitivity to the needs of local situations and individual students, an awareness of current curricular approaches, and the ability to apply these to a particular problem. Reports are important products in a variety of fields; laboratory and research reports and case studies in the field of social work are examples. The list could and should be further extended; for example, the evidence in certain management situations comprises memoranda, letters, and reports that constituted part of the way in which a situation was handled.

Product assessment can be used with both sponsored and nonsponsored learning experiences. With sponsored learning experiences, this method has the advantage of being planned in advance to demonstrate relevant skills, knowledge, and understandings in a fashion that can be evaluated fairly. Although certain elements of the product can be unique for each student, it is possible to stress the communalities, giving an opportunity to plan reliable and economical means of evaluation. Products of unsponsored prior learning experience do not have the advantages of planning, but they frequently have the advantage of being one of the few sources of external evidence available.

Some products of unsponsored learning experiences, such as work in the creative arts, may occur with such sufficient regularity that a working procedure for product assessment is worthwhile. The work of only a few students is evaluated at any one time, but over a period of time a relatively large number of students' products will have to be evaluated. In other cases, the products submitted will be diverse, and formal means of assessing products would take more time and effort than would be justified. Nevertheless, the ways in which these diverse, and sometimes, unique products, are evaluated can be influenced by the considerations, procedures, and models given in this chapter.

Product Assessment vs. Performance Assessment

Performance testing and product assessment need to be distinguished. In performance testing the student actually *performs* a given skill, such as playing a musical composition; in product assessment, all that is available is the end *product*, in this case, perhaps, a tape or a record of the student's performance. Whether to apply performance or product assessment depends upon a number of factors.

First, which is more important, the performance or the product? In the example given, the emphasis is on the performance. In the case of a musical composition, the product rather than the composer's particular method of composition embodies the important skills and understandings. However, even when the emphasis is on performance, a product may embody it in a useful way. A tape of a student playing a given composition can be regarded as a product, for example. This would allow a student to select the performance he considers his best, and might be appropriate as an indicator of prior learning. On the other hand, a tape made under structured test conditions with observers present might more properly be regarded as a performance test, and would be more appropriate if the ability to perform before a critical audience is part of the desired learning.

Second, practical considerations may determine the use of product assessment rather than performance testing. In some cases, the uses of products rather than performance may be more convenient, less expensive, and more reliable. For example, if several students make recordings of a musical composition (under the same conditions of recording), a jury can make comparisons and replay parts about which questions or differences of opinion arise. And, of course, in certain cases, especially in nonsponsored learning experiences, only the product is available.

Product Assessment and Written Material

Another distinction (or case of overlap) that may have to be made is between a written product and a written report or paper or examination. Obviously, if the purpose is to evaluate ability to write effectively, the methods for assessing free-response written material would apply. A paper written to demonstrate knowledge or understanding of a field would probably be evaluated as such papers usually are in academic situations. On the other hand, imaginative writing, such as plays, novels, poetry, can be judged as a product. But there are other cases in which the writing was produced for other reasons, naturally in the course of the experience, and is to be judged for the learning of other skills and understanding. Diaries and logs of experiences while traveling, case reports written by social workers, and laboratory and research reports are all examples of written products which can be assessed as evidence that specified skills and understandings have been learned. In many cases, such products could be used in conjunction with an interview. For example, a travel diary might implicitly reveal growth in understanding of another culture, but such growth would need to be checked explicitly in an interview.

An example of the use of a structured daily log is given in CAEL Institutional Report No. 3.[1] Students at UCLA in a number of courses in which experiences

[1] Jane Szutu Permaul and Marina Bühler Miko, *Documentation and Evaluation of Sponsored Experiential Learning*.

outside the classroom were incorporated kept daily logs in their field work or internships. In these logs they recorded both activities and reflections on what they thought they had learned. The logs were then analyzed by faculty experts for evidence of student learning, using a Process/Assessment Matrix. Five broad areas of learning—self-awareness, awareness of others, skill development, academic content, and career understanding—and five levels of involvement—identify, inform, describe, generalize, and apply decision-making skills—form the matrix. Thus, using the log as the product, the faculty experts could analyze it for evidence of student learning and of the level of that learning.

When product assessment is based on written material in any of the above senses, the problem frequently arises of distinguishing between the product and the writing. Much stress is placed in the academic world on well-written papers, despite the fact that student papers are frequently artificial. They are written by individuals knowing relatively little about a subject for other individuals who know much more. It is possible to place too much stress on this type of writing; poor products can be disguised by well-written reports and good products can be lost in poorly written papers.

A useful way of meeting this problem is to ask the question: What kind and level of writing is demanded if a particular product is to represent student learning of knowledge, understandings, and skills at a college level? At the one extreme, a novel or a series of poems must be judged on the quality of their writing. At the other extreme, the niceties of writing are not vital in a daily log. In between there will be a variety of possibilities. In a study of curriculum units judged as products, one of the best units presented contained no formal summary of what the student had learned from teaching the unit; instead each page of the unit was annotated with comments on what had and had not worked and how the unit should be changed before it was used again. The writing involved in products should suit the necessities of the situation in which the product is to be used rather than notions of academic propriety.

Product Assessment and Documentation

An important and necessary distinction needs to be made between product assessment and documentation. Students frequently present documents, such as letters from employers, newspaper clippings, licenses, and other materials, as evidence of experiences. Rarely do these afford evidence of learning, or if they do, of learning related to specific, relevant objectives. It is possible that a selection of documents produced on the job by a student could be the product equivalent of an in-basket test, but the skills, knowledge, and understandings to which they were related would have to be specified and evaluated. In most cases these products would need to be supplemented by an interview.

Product and Process

In many cases products cannot stand alone without some idea of the processes which produced them. When the product does stand alone, the burden is on the evaluator to identify the learnings represented. In the model of product assessment for the visual arts, three criteria have been identified, representing generalized

learnings in this area; but even in this case, a short paper or interview has been suggested to allow the student to clarify his intent in some of the work he presented. In the UCLA analysis of logs the matrix of criteria for learning has already been worked out and includes a process component.

While the product can testify clearly to some of the learnings, learnings associated with the process by which the product came into being need to be evaluated in other ways. For example, in evaluating ability to develop curricular material, many learnings are implicit in the unit itself—for example, ability to set objectives for students, to select materials, to evaluate changes in student behavior. Equally important process learnings are not explicit—for example, how the student identified the problem initially, how he gained the background needed to work on it, how he evaluates the success of what he attempted to do. Here both the product and the student's account of how he produced it are needed.

Developing Product Assessments

A brief sketch of the steps to be taken in developing a procedure for assessing products may be helpful for faculty members faced with the task of planning formal or informal procedures for evaluating learning from the evidence contained in products. The procedures suggested represent attempts to increase the validity and the reliability of the judgments to be made.

1. Identify Learning Outcomes. The first step is to identify the learnings expected of the student in the situation. Once these have been identified, perhaps with the student's help, it should be clear whether an assessment of a product will yield useful evidence of the quality of important learnings or will need to be supplemented by other methods of evaluation or will be inappropriate. A novel may stand by itself as evidence of the learning of skills in imaginative writing; the accounting records of a business venture may need to be supplemented by an explanatory paper or interview to identify and document further the relevant learnings; but what product will represent skills in interviewing—unless the student has taped several interviewing sessions?

2. Construct a Form for Product Assessment. The learnings already identified will constitute the criteria for the product assessment. Not only should the criteria be specified but also the relevant points to be observed in the product. If at all possible, levels of quality should be indicated in terms of characteristics of the product that are indicative of different levels of performance. For many products the best approach to defining criteria and products will be to provide specific examples of outstanding, average, and poor products.

A simple method is to construct a checklist, using each identified learning as one item on the list. For each item, a rating scale can be devised to judge the quality of the student's performance. The more clearly the rating scale is anchored in observable behaviors, the more consistently different judges can use it. At the time of construction, decisions can also be made whether certain learnings are to be given more weight than others in arriving at a summary evaluation.

The two models at the end of the chapter illustrate different ways in which formal assessment procedures can be set up; the Process/Assessment Matrix in

CAEL Institutional Report No. 3[2] gives still another possible variation. For informally assessing a unique or infrequent type of product, a simple list of learnings to be checked for the adequacy with which they have been achieved will help insure a consistent and a documented approach to product assessment.

3. Prepare Students. Guidelines need to be written for the student so that he or she understands the criteria by which his or her work will be judged. The student also needs instructions for preparing his product for evaluation, highlighting the important features of the product, and eliminating extraneous details which might adversely influence the evaluation. This step will usually not be difficult for sponsored learning experiences in which the product can be planned. Although more difficult for prior learning, such instructions are worthwhile even if they have to be modified in special cases.

4. Prepare Evaluators. Judges or evaluators also need guidelines for observing and judging products. In most respects these do not differ from the guidelines given students. The most important guideline for the evaluator is the form for product assessment.

If more than one evaluator is to be used, either to obtain more than one judgment of a product or to have several faculty available to work with students, training sessions using the evaluation method on a few sample products are useful to identify disagreements and permit discussions leading to revisions of the procedure and agreement on decisions.

5. Combining Product Assessment with an Interview or a Paper. Another part of planning the evaluation consists of deciding whether or not to combine the product assessment with an interview or a paper. Such a combination may be necessary if it is not immediately clear that the product gives evidence that the student has learned a certain skill. In addition, a product may reveal only some of the learnings possible in a given situation. There is the danger that ability to verbalize about the skills involved in a product may not be the same as the ability to make the product, and in the interview or paper some students may be rewarded for verbal skills rather than for those for which they are presumably being evaluated.

A MODEL FOR PRODUCT ASSESSMENT IN CURRICULUM DEVELOPMENT

To illustrate the possibilities of product assessment, two different plans for assessing products are presented. The first model is adapted from one developed for the master's program in education of the Juárez Lincoln Center of Antioch College. It was developed in a study of assessment procedures in the CAEL Operational Models Project.[3] This model consists of two documents, available to both students

---

[2]Permaul and Miko, *Documentation and Evaluation of Sponsored Experiential Learning.*

[3]See Ruth Churchill, Andre Guerrero, Janet Hartle, and Harry Horwitz, *Coordinating Educational Assessment Across College Centers.*

and faculty. The first is a statement of criterion standards for projects in curriculum development; the second, a checklist for evaluating such projects.

Program Statement in the Area of Curriculum

To meet minimum requirements the student must demonstrate a general knowledge of educational philosophy and its relationship to various curricular approaches. The student must be able to express this knowledge in terms of his/her own educational philosophy and the goals and/or objectives (state guidelines, local goals, etc.) to which the curriculum unit is related.

The development or revision of the curriculum unit is expected to include: general rationale, specific objectives, method of teaching and learning, data sources provided, and means of evaluating changes in pupil behavior produced by the curriculum unit.

In general, the student is expected to accompany the curriculum unit developed or revised with an account of how he or she identified the problem, acquired background for handling it, carried out the project, and evaluated his/her general learning from the project.

For major credit, a student must also teach at least part of the curriculum unit, provide a staff guide, evaluate the success of the curriculum unit in terms of pupil learning, and relate the project as a whole to other areas of professional development.

Checklist for Evaluating Projects in Curriculum Development

Name of Student _____ Name of Project _____

Educational philosophy and goals

1. Does the curricular unit reflect the student's educational philosophy?
   _____ Yes, clearly
   _____ Yes, in some respects
   _____ Questionable
   _____ No

   Evidence: (How does the student indicate his philosophy?)
   [Space left for free answer]

2. Does the project relate knowledge of educational philosophy to possible curricular approaches?
   _____ Yes, clearly
   _____ Yes, in some respects
   _____ Questionable
   _____ No

   Evidence: (What curricular approaches were explored? How is the curricular approach chosen related to educational philosophy?)
   [Space left for free answer]

3. To what goals, objectives, or guidelines has the curriculum unit been related?
   _____ State guidelines
   _____ Local goals and objectives (school or organization)
   _____ Teacher's goals and objectives
   _____ Pupils' goals and objectives

4. Has the curriculum unit been related to the goals, objectives, or guidelines chosen?
   _____ Yes, clearly
   _____ Yes, in some respects
   _____ Questionable
   _____ No

   Evidence: (Cite specific objectives and show relationship to curriculum unit.)

   [Space left for free answer]

Development or revision of curriculum unit

5. Is the curriculum unit clearly described and presented?
   _____ Yes, clearly
   _____ Yes, in some respects
   _____ Questionable
   _____ No

   Evidence: (What is the subject matter of the unit?)

   [Space left for free answer]

6. Is the rationale for the curriculum unit clearly stated?
   _____ Yes, clearly
   _____ Yes, in some respects
   _____ Questionable
   _____ No

   Evidence: (What is the rationale?)

   [Space left for free answer]

7. Are the objectives for the curriculum unit clearly stated?
   _____ Yes, clearly
   _____ Yes, in some respects
   _____ Questionable
   _____ No

   Evidence: (What are the objectives?)

   [Space left for free answer]

8. What is the method by which the curriculum unit will achieve its objectives?

   [Space left for free answer]

9. Is the method appropriate for achieving the given objectives?
   _____ Yes, clearly
   _____ Yes, in some respects
   _____ Questionable
   _____ No

   Evidence: (What methods are used; how are they related to the objectives?)

   [Space left for free answer]

10. What sources of data are provided in the curriculum unit?

    [Space left for free answer]

11. Are these data sources used adequately?
    _____ Yes, clearly
    _____ Yes, in some respects
    _____ Questionable
    _____ No

    Evidence: (Give an example of a data source and how it has been used in a unit.)

    [Space left for free answer]

12. By what means will the success of pupils in achieving the objectives of the curriculum unit be evaluated?

    [Space left for free answer]

13. Are the methods chosen suitable for evaluating this curricular unit?
    _____ Yes, clearly
    _____ Yes, in some respects
    _____ Questionable
    _____ No

    Evidence: (Give an example of a means of assessing what pupils have learned.)

    [Space left for free answer]

Use of the curriculum unit

14. Has the curricular unit been used?
    _____ Yes, the whole unit has been tried out
    _____ Yes, parts of the unit have been tried out
    _____ Questionable
    _____ No

    Evidence: (What has the student done to put the curricular unit into practice?

    [Space left for free answer]

15. Has a teacher or staff guide for this unit been prepared?
_____ Yes, one which can be used as it stands
_____ Yes, one which could be used with revisions and/or more work
_____ Questionable
_____ No

Evidence: (Could another teacher use the guide as it now stands?)
[Space left for free answer]

16. Has the use of the curricular unit been evaluated appropriately?
_____ Yes, clearly
_____ Yes, in some respects
_____ Questionable
_____ No

Evidence: (Cite means of evaluation, appropriateness, and adequacy of conclusions drawn.)
[Space left for free answer]

17. Is the area of curriculum development related to other areas such as development of sources of information, staff development, evaluation, community involvement, and bilingual-multicultural education?
_____ Yes, clearly
_____ Yes, in some respects
_____ Questionable
_____ No

Evidence: [Space left for free answer]

General

18. What sources of data have been used in carrying out this project?
[Space left for free answer]

19. Have these data sources proved adequate for the project?
_____ Yes, clearly
_____ Yes, in some respects
_____ Questionable
_____ No

Evidence: (Background in educational philosophy, curriculum development, specific area of curriculum development.)
[Space left for free answer]

20. Is the report written at an acceptable level for an educated adult?

_____ Publishable
_____ Adequate educated writing
_____ Needs work on writing
_____ Writing fails to communicate

Evidence: (Organization, clarity of presentation, use of words, grammar.)

[Space left for free answer]

21. As a whole, are you satisfied with this curriculum unit and the report on its development?

Major Credit*	Minor Credit*	
_____	_____	Clearly satisfactory
_____	_____	Adequate
_____	_____	Questionable
_____	_____	Not acceptable

*Pay no attention as to whether project was intended to receive minor or major credit. Rate it as to whether you believe it should receive major or minor credit. Major credit implies that all items have been carried out; for minor credit the curriculum need not have been used.

Evidence: (Outstanding strengths and weaknesses of the report.)

[Space left for free answer]

A MODEL FOR PRODUCT ASSESSMENT IN THE VISUAL ARTS

This plan is adapted from the CEEB Advanced Placement Test in Studio Art,[4] probably the best current example of product assessment based on research on judgments by art faculty of student work in art. It can be used as it stands, modified, or altered according to the purposes of the assessors. It can also serve as a model for further exploration of ways of assessing products of students.

Instructions to the Student

In order to earn college credit in the area of the visual arts—painting, graphics, sculpture, ceramics, photography, film, television, and other forms—you are asked to present for evaluation selections from the work that you have done. The purpose of the evaluation is to allow you to demonstrate your ability to deal with fundamental concerns in the visual arts.

The materials that you are asked to present deal with three objectives: overall quality of work, depth of work, and breadth of work. Evaluators are looking for works that reveal a style of execution and content initiated by you personally. The works presented may have been produced at any time on your own initiative or in art courses or other courses. The only restriction is that they should not be work for which you have already received college-level credit.

---

[4]College Entrance Examination Board, *Advanced Placement Course Description: Art.*

In the instructions that follow, certain instructions are given on the works you are to present and the ways in which you present them. The purpose of these instructions is to help you present a good sample of your work on the same basis as other students. In that way you can help the judges make a fair appraisal of your learning.

Overall Quality of Work. You are asked to present, *in their original form*, three to five works which you consider your best. These works will be judged for their overall quality. They may be individual works, or they may form a series. All of the works may be in one medium, or they may be in different media. The reason that you have been asked to submit a few of your best works is that experience indicates that large numbers of works include pieces of varying quality, and judges tend to evaluate the total group at a lower level than they would if a careful selection of a few works had been made.

- If the work, such as a film, was done in collaboration with other individuals, specify what were your responsibilities and contributions and what were those of others.
- Since these are original works and since they may be handled by several judges, try to keep the portfolio manageable in size and to protect the works from the consequences of unavoidable handling. Graphics that can be smeared should have a protective, transparent covering; small pieces, such as jewelry, should be securely mounted; tapes and film should be adequately packaged.
- Remember that the better you present your works, the less likely are the judges to be distracted by irrelevant considerations.

Depth of Work. To give you an opportunity to show your personal concentration on a particular way of thinking, working, and producing, you are asked to present evidence related either to an aspect of your work or to a project on which you have spent considerable time and interest and *which you may or may not consider successful.* Be sure that the material shows work on a particular activity or investigation carried out over a considerable period of time. Remember that the work should be one that you were personally committed to execute. Class problems in drawing or design, for example, are generally not acceptable.

Although the work may stand on its own merits and require no further clarification, you should present a short written or taped commentary to accompany the work. It helps the evaluators to know why you selected this area or project, what influences (things seen, heard, read, felt, imagined) affected your work, how planning affected the development of your work, and why you feel it is successful or not successful. [Alternatively: You will be asked in an interview to tell the evaluators why you selected this area or project, what influences (things seen, heard, read, felt, imagined) affected your work, how planning affected the development of your work, and why you feel it is successful or not.] If your work is part of a group project, your responsibilities should be clearly defined.

Most works for this section should not be submitted in their original form, with the exception of illustrated books and films, including videotapes. Sculpture, painting, murals, graphics, ceramics, jewelry, weaving, photography, and architecture may be illustrated by color slides. Environmental or conceptual projects can be described by slides, films, and/or appropriate written documents. Take care in

photographing your work. The quality of your slides has to be good so that judges can evaluate your work accurately. For three-dimensional work, provide several slides for each work, each slide showing a different view of the work. Try to keep your presentation short (no more than twenty slides or five minutes of film).

Breadth of Work. Here you are given an opportunity to show the range of problems, ideas, media, and approaches you have undertaken. You are asked to submit slides of completed works that show how you have solved problems in the following categories:

- spatial illusion on a flat surface
- color
- drawing
- organization of three-dimensional figures

On the slides identify the problem you are attempting to solve, for example, use of space or color relationships. This section is one in which evaluators look at aspects of your work in color, design, and technique relatively isolated from the total quality of the work or from what you wished to express (expressive intent). Total quality always influences evaluators' judgments, but art skills are emphasized in this section.

Since any work submitted will probably incorporate more than one category, use works that clearly emphasize one category even though other problems may be present. Works appearing in this section may already have appeared in one of the first two sections. Make a careful selection of slides; try to keep your presentation short (no more than twenty slides).

Instructions to the Evaluators

Works of art are highly personal and subjective in content and expressive means. They can only be evaluated if judges are firmly committed to this concept. The most important part of the assessment is the student's response to *Overall Quality of Work*. Quality is also an important part of the judgments made in *Depth of Work* and *Breadth of Work*.

Clearly the student must perform at college level if he or she is to receive credit. Since professors make judgments as to whether or not their students are performing satisfactorily in college courses in art, it is assumed that they can judge college-level performance on the basis of a carefully selected portfolio of art works submitted by students whose work in art is based on either sponsored (current and planned) or nonsponsored (prior and nonplanned) experiences.

The validity of the three criteria proposed (overall quality, depth, and breadth) is based on research on the factors which college instructors in art schools have found important and about which they could agree in their judgments. Validity of the judgments is also affected by the selection of the judges. Generally judges will be from the institution's own art faculty, but some institutions will want to include practicing artists and art faculty from other institutions. It is important that the judges review the criteria and make sure that they agree with them. An art faculty may want to formulate its own criteria, both in terms of how they state them and what works they ask students to submit. In that case, the procedures set forth here can be used as a model and adapted.

At least two judges should be used in the product assessment of studio art. Not only does the use of more than one judge increase the validity of the judgments because of the broader background in art represented, but it also increases the reliability of the judgments by cancelling out individual idiosyncracies. Reliability can also be increased by a training session. Two or three sample folders of student work can be made up, and each evaluator can judge them independently. After the independent judgments have been made, individual agreements and disagreements of the judges can be discussed, and the judges can arrive at an agreement on their ratings. This procedure gives the judges some assurance that they are looking for the same characteristics and using the same scores for works of similar quality. Total agreement of judges is not sought but rather a general tendency to agree.

In making the evaluations, the judges should take certain general precautions to insure that all students are judged under the same conditions as far as possible:

1. Judges should work independently. No judge should know another's scores until the judges have finished.
2. The work of judges should be scheduled in advance. Judges should spend approximately the same length of time; reasonable minimum and maximum lengths of time should be worked out. Necessary equipment should be available so that time is not wasted.
3. If an interview is to be included, judges should view the works in the twenty-four hours preceding the interview.

Rating of Overall Quality of Work. A global judgment of quality based on three to five original works of art is called for. The most important factors to look for are:

- Personal expressive content of the work.
- Appropriateness of the student's choice and execution to the medium.
- Unique solutions.
- Avoidance of tricks or obvious artistic mannerisms.

Faculty members engaged in evaluating student work will probably want to specify factors to look for in the works that are indicative of level of performance in this section and the other two sections. In this way faculty understanding of the levels can be clarified and greater agreement in ratings achieved.

For rating, a general four-point scale is suggested for use here and in the other two sections:

1—If the original works suggest that the student is decidedly not working at a college level.
2—If the original works suggest doubtful college material.
3—If the original works are equivalent to those of a college student beginning to work in art.
4—If the original works can be considered equivalent to the work of a college student doing advanced work in art.

Rating of Depth of Work. Here works or slides of art works reflecting a long-term effort on the part of the student to solve a problem in art are judged. Usually the

student submits up to twenty slides of work, although films, videotapes, and illustrated books may also be submitted. The project should be accompanied by a short commentary telling why the student selected the area or project, what influenced the work, how planning affected the development of the work, and whether the student feels the work was successful or not. (Alternatively, the judges may wish to interview the student on these points.)

Important factors to be considered in judging are:

- Concentration on a particular project or problem over an extended period of time.
- Personal commitment to the content and execution of the project or problem.

Rating of Breadth of Work. This section calls for a judgment on several aspects of the student's work in color, design, and technique relatively isolated from the total quality of the work. Students submit slides dealing with: spatial illusion, color, drawing, and organization of three-dimensional forms.

Overall Rating. It is necessary to combine the ratings on the three sections of the product assessment into an overall rating, expressing the level at which the student's work has met the criteria set. The first decision is whether or not to weight the three sections of the assessment equally. Some may wish to weight overall quality twice as important as the other sections. For each judge an overall average should be computed, and then the overall averages of the judges should be averaged together to give a single final rating.

It is recommended that final ratings above 3.0 be awarded credit or recognition at the level of a college student working in art at an advanced level and that final ratings above 2.0 and below 3.0 be awarded credit or recognition at a beginning level for college students. These recommendations recognize the fact that each judge has made three separate judgments, one for each section, and that a student who submits work high in all three respects is performing better on the whole than the one who has met a satisfactory level in only one or two sections. The amount of credit or recognition granted will differ from college to college, depending on college and departmental policies.

Bibliography On Product Assessment

Churchill, Ruth; Guerrero, Andre; Hartle, Janet; and Horwitz, Harry. *Coordinating educational assessment across college centers.* CAEL Institutional Report, Antioch College. Princeton, N.J.: CAEL Educational Testing Service, 1976.

College Entrance Examination Board. *1975-76 Advanced Placement course description: Art.* New York: CEEB, 1975.

Dorn, Charles M. *Evaluating the Advanced Placement studio art portfolio.* Princeton, N.J.: Advanced Placement Program, College Entrance Examination Board, 1974.

Dyer, Henry S. College testing and the arts, in Lawrence E. Dennis and Renate M. Jacob (Eds.), *The arts in higher education.* San Francisco: Jossey-Bass, 1968, pp. 82-104.

Fitzpatrick, Robert and Morrison, Edward J. Performance and product evaluation, Chapter 9 in Robert L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 1971, pp. 237-270.

Permaul, Jane Szutu, and Miko, Marina Bühler. *Documentation and evaluation of sponsored experiential learning.* CAEL Institutional Report No. 3 (revised ed.). Princeton, N.J.: CAEL, Educational Testing Service, 1977.

V. Performance Assessment

Arnold Fletcher and John L. D. Clark

NATURE AND BASIC ISSUES

The performance test is usually described as a test that evaluates a person's ability to carry out or execute an action, process, or task. Although paper-and-pencil items that involve the writing of symbols in response to symbolic stimuli may be used, performance testing more characteristically stresses the ability to manipulate something or perform a role—for example, typing, piloting a plane, assembling a piece of machinery, playing the violin, teaching, acting, or leading discussions.

Some Distinguishing Characteristics of Performance Testing

In attempting to place performance testing in an assessment perspective, we must emphasize that much, but not all, of such testing involves the measurement of skills, described in dictionary terms as "developed or acquired abilities," which are usually identified with psychomotor activities. In addition, some educators tend to treat skill testing, performance testing, and competence testing interchangeably. Although related, they are not necessarily synonymous; using them synonymously can cause confusion about the concept of performance testing and can lead to the misapplication of testing techniques.

Kelley describes competence-based education in terms of four categories of objectives: (1) *cognitive*—what the student should know; (2) *affective*—beliefs, values, and attitudes of the individual; (3) *performance*—specified behavior to be demonstrated by the learner; and (4) *consequence*—student learning identified by the teacher.[1] Thus, a skill test tends to evaluate a more specific limited ability usually within a larger performance context, whereas a performance test generally attempts to assess a higher level, more complex behavior, which, in turn, is often part of an even more comprehensive competence.

Another distinguishing characteristic of the performance test is the presence of a large element of realism, or what is sometimes described as situational realism. The student is expected to perform in as close to a real-life situation as possible. Since it is very often impractical or impossible to set up such a situation, the assessor must simulate reality. This practice is characteristic of student-teacher training programs, the type and degree of simulation being dependent upon the particular needs of the student. Fitzpatrick and Morrison indicate that for a simulated situation to be representative of a real-life environment it must (a) be *comprehensive*—simulate a range of varying aspects of the situation, and (b) demonstrate *fidelity*—reflect a desirable degree of reality.[2]

As a final note, it may be useful to reiterate the differences between performance and product assessment. As pointed out in the previous chapter, perfor-

---

[1]Edgar Kelley, *Three Views of Competency-based Teacher Education*, pp. 9-10.

[2]Robert Fitzpatrick and Edward J. Morrison, "Performance and Product Evaluation," p. 240.

mance assessment often results in a product that can also be assessed. Where an end product is available, the emphasis in performance assessment is on the antecedent steps performed to arrive at the product, rather than the product itself. Systematic procedures for observing and recording behavior are usually characteristic of performance assessment whether or not a product emerges. In addition, performance assessment procedures are usually characterized by externally structured conditions that remain constant for different individuals whose performance is being assessed.

Types of Tests

Classification in any field poses the problem of overlap. Tests vary so much in terms of purpose, content, and procedure that different authorities approach classification in different ways. The categories, descriptions, and examples presented herein are not intended to be all-inclusive but, rather representative of the principal types in current use. The reader is also referred to the appropriate works in the bibliography for additional treatment, particularly the CAEL publication *A Compendium of Assessment Techniques* by Knapp and Sharon.[3]

The *work sample* is one of the most common types of performance testing, usually possessing elements of simulation. In this case, job-related tasks are set up, following which the student or trainee is evaluated for the task performed or the product created. The work sample tends to recreate an important part, if not all, of the actual job tasks and operations, frequently using job equipment. The test will sometimes involve certain on-the-job difficulties, which the examinee must overcome. Although some of the realism may be sacrificed, the critical job elements are present and the test is readily recognized as exemplifying situational realism. Referral to other CAEL publications, prepared for both faculty and student use, is suggested for further coverage of work sample assessment problems.[4]

The *situational test* endeavors to evaluate performance in a real-life or simulated situation in which the examinee is asked to undertake a task or role that has been thoroughly explained before being carried out and assessed. Such tests usually attempt to evaluate more complex performance qualities, oftentimes in situations involving human interaction. Some representative examples are the following:

1. War games and business games.
2. Role-playing in supervision and management, criminal justice, and social work.
3. In-basket techniques in administration and office practice.[5]
4. Employer evaluation of employee performance.
5. Case studies in sociology and human services.
6. Simulated clinical problem solving in allied health fields.[6]

---

[3]See especially pp. 7-23.

[4]See, for example, *Teaching and Assessing Interpersonal Competence–A CAEL Handbook* and *Learning and Assessing Interpersonal Competence–A CAEL Student Guide*, by Paul Breen, Thomas F. Donlon, and Urban Whitaker; and *Assessing Prior Learning–A CAEL Handbook*, by Joan Knapp.

[5]Lois A. Crooks, *Issues in the Development and Validation of In-basket Exercises for Specific Objectives*, pp. 1-10.

[6]Christine H. McGuire and David Babbott, "Simulation Technique in the Measurement of Problem-solving Skills," pp. 1-10.

7. Laboratory procedures in industrial research.
8. Simulation of field projects in training for Peace Corps service.
9. Student teaching in teacher education.

Simulator assessment typically involves a machine, piece of equipment, or material designed to simulate parts of an operational situation in which a person's performance is to be evaluated. According to Gagné, the simulator is characterized by three features: (1) it is representative of a real situation; (2) it is designed to provide specific controls over the situation; and (3) it deliberately omits some aspects of the operation that are variable or unpredictable.[7]

There are a great variety of simulators ranging from the simplest equipment, such as the artificial rifle for individual drill in rifle handling, to the complex systems simulator designed for use by whole teams of military personnel. The relative degree of complexity depends on the task or operation being evaluated. Gagné classifies such operations as follows:

1. *Procedures* such as cockpit training for pilots, console drills to control missile firing, or procedures for calibrating electronic gear.
2. *Motor skills* as in simulated gunnery practice or aircraft landing.
3. *Identifications* stressing immediate recognition of targets, types of airplanes, or terrain features.
4. *Conceptual tasks* related to equipment testing, navigation, and interpretation of aerial photos.
5. *Team functions* as in complex war games or communications among crew members.

The *assessment of prepared performances,* another form of performance testing, stresses the evaluation of artistic renditions or demonstrations of physical prowess that are normally the result of repetitive rehearsing for a period prior to the evaluation. Typical examples include rehearsed performances in the performing arts—music, dance, and theater arts—and in sports and physical education activities.

In some instances, the test will be devised to evaluate purely individual efforts as in solo vocal and instrumental auditions, forensics, gymnastic exhibitions, track and field events, and physical fitness programs. In others, the evaluation of individual performance must be given in the context of group activities, a factor which poses many assessment difficulties such as the ability to isolate the performance and control the testing situation. There are surprisingly few standardized tests in the performing arts fields because the major thrust has been to develop measures of artistic aptitude, appreciation, and aesthetic attitude. In music, the best example of a test designed to assess both rehearsed and unrehearsed music is the *Watkins-Farnum Performance Scale,* a standardized test for band instruments developed originally in 1942 but still a significant contribution to music performance testing.[8]

On the other hand, the field of physical education has given a great deal of attention to the development of tests that evaluate the results of previous practice.

---

[7]Robert M. Gagné, "Simulators," pp. 225-226.

[8]John G. Watkins, *Objective Measurement of Instrumental Performance.*

Tests range from those measuring specific, discrete skills in individual sports and physical fitness to more gross measures of performing proficiency in either individual or group situations. A comprehensive, useful treatment of this complex area of measurement is found in the recent textbook of Franks and Deutsch, *Evaluating Performance in Physical Education,* which proposes a hierarchical model of physical performance and fitness as a basis for assessment, the major components of which are body composition, efficiency, endurance, skills, and sociopsychological influences.

Unique Problems in Performance Testing

All assessment operates on the basis of a common body of principles concerning which psychometricians are in reasonably uniform agreement. On the other hand, the problems that will arise in the application of these principles will vary considerably in kind and degree depending on the nature of the assessment situation. Thus, the utilization of expert judgment in the performance test poses different, though not necessarily more severe, assessment problems than it does, say, in the oral interview often used as a measure of subject-matter knowledge.

One of the most difficult and unique problems to be faced in assessing performance is the question of its representativeness and its relationship to the twin measurement demands of validity and reliability. Although these concepts were discussed in Chapter II, the uniqueness of the problem requires a separate treatment here.

Simulation and the Validity Question. Earlier, it was pointed out that the principal characteristic of the performance test is its relative realism or simulation of a real-life situation. Furthermore, a simulation is considered to be representative of a real situation when it is *comprehensive,* i.e., covers a wide range of aspects, and when it demonstrates *fidelity,* i.e., faithfully reproduces each aspect. When these two conditions are adequately met in a performance test situation, one can assume that the test has achieved a satisfactory degree of validity. Stated in assessment terms, a performance test should produce a more valid result when the elements of the criterion situation are adequately sampled and realistically represented. The failure to simulate adequately a real situation is one of the chief shortcomings of a great many applications of performance assessment. A case in point has been the common practice in teacher education of assessing the ability of college students to understand principles of teaching on the basis of an observation of their teaching performance in a simulated classroom environment where their peers pretend to be young children.

To illustrate more fully the need for adequate simulation to insure test validity, let us consider the case of a final examination administered at the close of the training period for a student preparing to become a stenographic reporter. In this example, assume that the basic manipulative and transcriptive skills have previously been tested and that the objective of this examination is to assess the ability of the student to perform satisfactorily in a simulated on-the-job situation, in this case represented by an audiotape of a courtroom scene.

To recapitulate from the previous discussion, the simulation must produce conditions that satisfactorily answer the following questions:

1. Does the test create sufficient situational realism?
2. Does it provide an adequate sampling of the various elements in a courtroom scene that are crucial to the performance of the stenographer?

The following checklist of preparations is suggested to provide the simulation necessary to produce valid test results:

1. Elements of the courtroom scene.
   a. Prepare two tapes of the scene, a first version for tryout with students and a second refined version.
   b. Set up a scene that utilizes a number of voices of different qualities, accents, and inflections.
   c. Provide a combination of longer single speeches and rapid-fire verbal exchange involving several persons.
   d. Have the dialogues include a mix of declaratory statements, questions, short, long, and incomplete sentences, and interjections in the middle of statements and questions.
   e. Create subdued interchanges, as between judge and counsel, and speech at high-volume levels.
   f. Provide for background masking sounds such as spectator laughter, rustling of papers, and the like.
2. Equipment considerations.
   a. Use recording and playback equipment of good fidelity, particularly for reproduction of the spoken voice which, to be intelligible, depends on recording and reproduction emphasizing higher frequency responses.
   b. Use loudspeakers, preferably two (front and back), rather than earphones to preserve the sense of sound coming from several directions.
   c. To preserve sound quality store the original tape and use copies for frequent repetitions.
   d. Make sure stenographic equipment is in good working order.

One method of handling the validity problem in simulations might be for the assessor to begin with a highly simulated testing model and, under successive tryouts of the test situation, remove elements and note the effects. For example, in the illustration presented above, it might be feasible to create a videotape recording of the scene first, extract an audiotape version, and compare the test results achieved from both versions. At the close of this process, the assessor could well find that, for some assessment situations, a less closely simulated model would serve to satisfy validity needs.

Relationship to Reliability. The validity problem has direct bearing on the reliability problem and, in fact, can pose a real dilemma for the assessor. Real-life conditions, particularly where complex human interaction situations are concerned, are very difficult to control. Now, it is a *sine qua non* that sound measurement demands control; the examinee should be evaluated each time under similar conditions. Only then can we expect a test to be reliable, i.e., consistently produce the

same result over a period of time. By increasing the representativeness of the simulation, we assume an increase in validity, but therein lies the dilemma. As validity increases, the possibility for control generally decreases and so, therefore, does reliability. Without reliability, of course, validity has little meaning. Since both elements must be present, the assessor has no choice but to create a reasonable balance between the two in simulating appropriate conditions for performance testing, an achievement that can be realized only after experimentation over an extended period of time.

Another way of looking at the problem of reliability is to picture the evaluation of performance in educational settings as an assessment continuum. At one end we would be dealing with one-on-one evaluation, one expert judging a single performance, and at the other large numbers of performance evaluations utilizing a standardized test, such as the *Watkins-Farnum Performance Scale* designed to assess performance on band instruments. In the one-on-one situation, little disagreement would be likely to emerge among experts when evaluating lower levels of performance, for example, the assessment of beginning typing or shorthand. In situations of this kind, the use of a single evaluator would probably be justified. But as we move up to higher, more complex levels of performance where the variables increase, as in evaluating artistic renditions or performance in leaderless group discussion, and the value-oriented judgments of assessors become a contending factor, the likelihood of disagreement is much greater. In such situations, the ability to achieve an acceptable degree of reliability becomes more difficult when using the single expert. At the other extreme, the use of the standardized performance tests in institutional settings can be completely unfeasible. Although many such tests report good reliabilities, the low volume of students creates a very high cost factor. In addition, this type of test is usually so inflexible as to be inappropriate for most performance situations.

A reasonable solution for the reliability dilemma lies somewhere between these two extremes. The compromise normally would call for the use of multiple ratings, generally involving the so-called jury system of assessment. The more institutions become involved in competence-based education, work experience programs, community-oriented activities programs, and the recognition of lifelong learning accomplishments, the greater becomes the need to develop performance assessments in the jury mode. Although the cost may still be a factor for institutions with small faculties, the trade-off for improved assessment could be great. In utilizing this assessment mode, the addition of a crucial element should be given serious consideration—the process should be supervised by institution-based psychometric specialists assigned to teams of subject specialists functioning in fields where performance is demanded. An example of a model of this team approach to performance assessment is presented later in this chapter (see p. 71).

As an example of the validity-reliability problem, the evaluation of written performance in a foreign language presents interesting challenges from the point of view of both test validity and reliability. A high, face-valid technique in this regard would be to have the student prepare one or more real-life documents such as memoranda, personal or business letters, and so forth, and to have these evaluated for overall quality. Although the validity of such a testing situation would be self-evident, the reliability with which the students' performances could be scored might be called into question since global appraisals of "quality"—

especially of fairly lengthy and largely undirected texts—can produce wide variations in the scores assigned, depending on the particular individuals doing the evaluations, the amount of preliminary training in the scoring techniques, the sequence in which the texts are read, and a number of other factors extraneous to the question of writing proficiency per se.

To provide for a greater degree of reliability in the writing test situation—especially when the papers must be graded individually by a number of different instructors—one possible approach would be to break the student's writing task into a number of smaller and more highly controlled elements, each of which could be evaluated more precisely and more uniformly than would be the case with a longer and "freer" text. Formats in which the student is asked to complete partial sentences by filling in appropriate verb or adjective forms, or to produce whole sentences which incorporate certain specified elements, usually permit easier and appreciably more reliable scoring.

These and other types of writing tests involving student performance on a series of discrete elements may, of course, be viewed as somewhat less "genuine" measures of writing ability than is the production of an entire business letter or other texts typical of real-life writing activities. If the former kinds of exercises are considered insufficiently representative of the types of writing tasks that the student will, in fact, encounter outside of class, it may be necessary to adopt an intermediate approach such as the writing of short portions of each of several representative writing tasks (for example, the opening paragraph of a business letter, with the topic and purpose of the letter specified in advance). This procedure would have somewhat greater situational realism than the discrete element task and would at the same time permit more reliable scoring than is the case with longer complete themes.

Checklist of Considerations

In concluding the treatment of this phase of assessment, the potential evaluator should be alerted to the special conditions that ought to be satisifed in developing and administering performance assessments. Considerations specific to performance assessment can be derived from the more general steps presented at the end of Chapter II.

The method of presentation is a checklist of five categories of key questions to be considered.

1. *The performance objectives*
   a. What are the crucial behaviors involved in the performance? Are some more important than others, and should they be weighted?
   b. Is sufficient information available, or will detailed task analysis be necessary? *Example:* The more readily identifiable tasks of the typist as compared with the complex range of the functions of a supervisor of 25 file clerks, typists, and stenographers.
   c. Under what specific conditions are the behaviors elicited in real life?
   d. Are appropriate measurement instruments already available, thereby making it unnecessary to "reinvent the wheel"?

2. *The choice of assessment method*
   a. Can a real-life situation be utilized, thereby avoiding the necessity of simulation? If so, can controls be set up to satisfy reliability needs?
   b. What is the cost factor to consider in determining whether to use a real situation?
   c. If simulation is necessary, what are the crucial aspects to be simulated? Is equipment adequate? Physical arrangements proper? Sufficient displays and aids?
   d. Is the assessment set up to enable close replication for evaluating different students as well as the same student at different times?
   e. Which is more appropriate, product or performance assessment, or both?

3. *The assessment process*
   a. Have appropriate criteria been set up to choose assessors properly?
   b. Have adequate plans been made for proper observation and recording of performance? *Example:* Determination of the appropriate means of recording among tape recorder, videotape recorder, or movie camera.
   c. Have necessary scoring procedures and forms been set up, such as checklists, rating scales, etc.?
   d. Have absolute and relative weights of assessment elements been established? *Example:* Evaluating the pilot in handling essential landing procedures versus the form and smoothness of the landing.
   e. Can one expert do the evaluating, or are multiple raters preferable? *Example:* Generally sufficient to use single rater in evaluating lower level skills; advanced levels require several raters.
   f. What training procedures have been developed for assessors?
   g. Have scoring standards been checked periodically when tests are repeated over a period of time?

4. *Assessment administration*
   a. Have instructions for the student been developed?
   b. Have necessary assessment controls been established, including desirable physical conditions, sequence and timing of activities, and the like?
   c. Has test security been worked out, such as the development of sufficient alternate lists of tasks from which to choose?

5. *Economic considerations*
   a. Are the results of the test of such importance that a high cost may well be justified? *Example:* The necessity of evaluating a real performance of the potential conductor of a professional symphony orchestra as compared with accepting the simulated testing of the student who must satisfy conducting requirements for the Bachelor of Music degree.
   b. How long is the assessment or test expected to be viable?
      (1) Will the volume be such that developmental costs can be recouped over a period of time?
      (2) Is the field changing so fast that the particular test will be outmoded?

c. What are the future needs for modification of the assessment system? That is, how much flexibility will be necessary regarding performance content or tasks and standards?

A PLAN FOR THE USE OF EXPERT JUDGMENT IN PERFORMANCE ASSESSMENT

This section presents an example of a procedural plan for performance-based assessment of experiential learning. It was developed at an external degree institution that relies for the development and administration of its assessment procedures on a close working relationship between the institution's professional staff and a group of approximately 300 outside academic specialists. The plan is viewed as a procedural goal, although most of its components are already in use. Other institutions may find it useful as a model of one possible approach to performance assessment. The description of the plan is followed by representative illustrations of performance assessment.

Selection and Orientation of Performance Assessors

Assessment staff at the college or university solicit recommendations of qualified assessors from academic deans and other college or non-college sources. Important criteria are: subject-matter expertise, an appreciation of the operation and philosophy of external degree programs, and a willingness to become familiar with performance-based techniques for measuring student accomplishment. Vitas of proposed assessors, lists of courses taught, and other documents are reviewed as part of the selection process, and wherever possible an informal interview with the prospective assessor is carried out by college staff. An ongoing informational file is maintained for each participating assessor showing the assessment projects in which he or she has been involved and related comments of the assessment staff of the institution.

Each newly-appointed assessor receives a package of explanatory materials describing the organization and philosophy of the college and the basic procedural steps to be followed in the assessment program. (A loose-leaf handbook, which can be added to and revised on the basis of continuing experience, is a useful aid.)

Following their review of background materials, assessors attend a joint discussion meeting with the college assessment staff to receive more specific information on the particular type of project in which they will be involved and to discuss the general outline of the assessment plan. To the extent possible, all assessors for a given performance area are asked to be present simultaneously, so that all participants receive the same type of briefing and orientation to the project.

Development and Tryout of Assessment Procedures

In situations involving on-site performance evaluations, assessors and college staff make an exploratory visit to determine that the physical layout and "ambiance" of the site, including the presence and condition of necessary equipment or other support elements, permit a highly valid, "real-life" demonstration of the performance to be assessed.

One or more assessors, aided by college staff as necessary, review published instruments or documented procedures in the area to determine if appropriate measurement tools are available or can be readily adapted from existing sources. Any findings are reported to the full group of assessors for discussion.

Assessors and college staff draw up a formal plan for eliciting, observing, and evaluating the student performance. The primary considerations in the plan are:

a. Providing for an adequate and representative sampling of the performance to be assessed.
b. Presenting directions to the student that adequately orient and prepare him or her for the task. These can range from a short list of procedures to well-developed study guides.
c. Developing an observation system, including any necessary forms, that can be efficiently and accurately used in the actual assessment situation.

Assessors and college staff, working in close collaboration, develop necessary procedures and instruments to implement the evaluation plan. A variety of approaches are used, depending on the number of assessors working on a given project and the relative difficulty of conceptualizing and/or developing appropriate assessment techniques. In very straightforward subject fields, where all assessors are quite clear about and in agreement on the proper evaluation techniques, individual assessors may be assigned to work up different specified portions of the entire assessment procedure, which are then combined, often with slight revisions following group discussion. In more complex situations, the assessors may work as a group, with a high level of input by the college assessment staff, and may also submit their work for critique and suggestions by an additional consultant in the specialization.

Simultaneously with the development of procedures and instruments, the assessors and college staff discuss the criteria and standards to be used in judging the student performance. Depending on the nature of the field, the student may be required to meet certain minimum performance standards in each of several subareas, or superior performance in one aspect of the assessment may legitimately be considered to compensate for less successful performance in other aspects. Related decisions are made on whether a simple credit no-credit evaluation report will be made or whether several different levels of accomplishment will be identified and reported. Additionally, a decision is made on whether supplementary narrative descriptions of observed strengths and weaknesses in the student's performance will be provided. If necessary, assessment plans and draft instruments are revised to meet the information-gathering and reporting goals decided on.

In situations where the volume of assessments warrants it, a preliminary tryout of the assessment procedures will normally be conducted, in which each of the assessors for that project simultaneously evaluates a number of student performances by using the draft procedures developed. Results are reviewed in terms of the efficiency and straightforwardness of the rating procedure and the extent to which the assessors are able to give closely similar ratings of any given student. Problems in either of these two areas indicate the need to discuss and revise the assessment procedures.

81

Operational Application of Assessment Procedures

Each student's performance is evaluated simultaneously by at least two assessors. As provided for in the assessment plan, the student has been fully alerted and "prepared" for the assessment process and is given the opportunity to present the total range of behaviors under consideration. In some instances, this requires observation of the student's performance on two or more separate occasions. On each occasion, the student is granted a warm-up period to become acclimatized to the situation, including any equipment or apparatus.

Although the assessors are present simultaneously at the performance(s) of the student, rating notations are made independently for later discussion and reliability checking. (Note: In some performance assessment situations, it is impossible for financial or other reasons to use more than a single assessor per student. In these instances, initial *group* development of the assessment procedure and group determination that the rating technique does permit reliable scoring by a single individual offers some justification for the single-assessor approach. In ongoing programs that must use a single-assessor technique, periodic spot checks, in which one or more additional assessors are called in to give independent ratings of the student's performance for comparison purposes, are carried out wherever possible.)

Scoring and Reporting of Student Performance

When more than one rater is used for the evaluation of single or multiple performances, the assessors and college staff meet as a group to review and discuss the student performance ratings. Tabulations of these ratings and related statistical descriptions are presented and interpreted to the assessors by college staff.

Average ratings of all assessors evaluating a given student constitute the "official" rating, except when an obviously aberrant rating is discounted through group discussion and agreement. Narrative descriptions of performance, including any applicable suggestions to the student regarding needed additional work or preparation, are developed on a group basis and made part of the final report to the student.

Project Review by College Staff and Assessors

A portion of the scoring and reporting is devoted to a critical analysis of the entire assessment project, including suggestions for conceptual and procedural improvements that could be made in future repetitions of the process, either within that subject-matter area or in the college's assessment program generally. Internal staff review of each completed assessment project is also carried out for the same revision and development purposes.

ILLUSTRATIONS OF PERFORMANCE ASSESSMENT[9]

1. Classroom Performance of Child Care Center Trainees

Special Assessment Techniques

- Quarterly observations of teaching
- Use of checklist to identify 14 competences expected of each trainee
- Use of two assessors (on-site trainer and outside consultant) to provide independent ratings
- Structured interview with each trainee based on individual anecdotal record and interview guide
- Examination of individual assignments

2. Piano Performance

Special Assessment Techniques

- Live and taped performance covering scales and arpeggios, three compositions from classic, romantic, and contemporary periods, and sight reading
- First assessor evaluates live performance using rating scales covering various elements of performance, five-point scale, average of three or more as passing
- Taped performance evaluated by second assessor and ratings compared

Comments. The second assessor should be provided with copies of the relevant music, including the music used for the sightreading. The credit allocation recommendation is controlled by the first assessor.

3. Painting and Drawing Competence in Studio Art

Special Assessment Techniques

- Examination of prepared portfolio containing examples of work in several media such as water color, oil, gouache
- Observation in studio setting of drawing skill
- Response to set problems requiring application to demonstrate basic drawing and composition skills

Comments. A second assessor can concentrate on portfolio evaluation, which should provide evidence of the applicant's competence. Both raters should examine material looking for evidence in growth and development of painting and drawing skills.

4. Alcoholism Counseling

Special Assessment Techniques

- Periodic observation by means of a one-way mirror and/or videotape of actual counseling situations
- Participation/observation in group sessions
- Personal interview to determine applicant's self-awareness of problems of addiction

---

[9]The first eight illustrations from the files of Thomas A. Edison College of New Jersey. The technique described in the last illustration on foreign language speaking proficiency was developed as a part of the Language Proficiency Program of the Higher Education and Career Programs Division, Educational Testing Service, Princeton, New Jersey.

Comments. Assessment should incorporate standardized means to test the applicant's basic knowledge of sociological, psychological, chemical, and physiological bases of addiction.

5. Flight Skills of Individuals Already Holding FAA (Federal Aviation Administration) Ratings

Special Assessment Techniques

- Observation by a certified FAA examiner of an actual performance
- Codified standards used to rate performance are those developed by FAA panel of experts that set licensure requirements

Comments. The format of assessment and credit practices follows recommendations of the AERO (Aviation Education Review Organization) College Credit Standard Guide. Assessment is available to individuals holding ratings for Private Pilot, Commercial Pilot, Instrument Pilot, Flight Instructor, and Instructor Flight Instructor.

6. Medical Laboratory Technology

Special Assessment Techniques

- Observation in a clinical setting of performance of listed laboratory techniques
- Performance of a set of laboratory tests on prepared samples in areas of hematology, urinalysis, serology, coagulation, immunohematology, clinical chemistry, and bacteriology

Comments. Knowledge of laboratory techniques may or may not be coupled with academic knowledge of related areas. Such knowledge can be tested through standardized means and credit awarded in natural sciences or as free electives where appropriate. College-level academic knowledge is not a prerequisite for ability to perform in a clinical setting.

7. Physics

Special Assesment Techniques

- Selection, on a random basis, of laboratory problems that simulate actual problems
- Designation of problems that can be solved with "pencil-and-paper," problems that have some relationship to real-life problems
- Observation of problem-solving ability in both laboratory and classroom type setting
- Review of laboratory reports prepared according to standard methods to determine that this aspect of the performance is adequate
- Second assessor to examine laboratory notes and responses to "pencil-and-paper" problems

Comments. Performance in physics often requires the working out of relationships for the purpose of solving a problem. This applies to both the classroom and work place. The assessment should be designed to evaluate the particular competences covered in a college curriculum.

In this kind of assessment, it is often possible to evaluate for advanced knowledge and competence and verify elementary-level knowledge at the same time. Specific aspects of that knowledge not covered by the higher-level evaluation can be evaluated in the form of additional problems.

8. Home Economics—Food Preparation

Special Assessment Techniques

- Definition of broad-scale problem that will involve the evaluation of nutrition and other dietetic factors related to meal preparation
- Designation of a specific situation that will call on student to demonstrate a variety of competences and the knowledge requisite to a satisfactory outcome
- Preparation of a "shopping list" that will adequately cover the needs stated above
- Evaluation of items purchased
- Observation in a kitchen setting of actual meal preparation
- Sampling of outcome and evaluation of final product

Comments. Various stated criteria will have to be met. These would be the standard agreed-on criteria that should be accepted by anyone in the field. The particular method allows for the statement of a problem whose solution calls on a variety of skills at different levels and satisfies the demands of different parts of the curriculum. One or more of the following examples of parts of the curriculum can be stressed: volume feeding, gourmet cooking, cooking for the invalid, nutrition, consumer education, product development in a commercial setting.

9. Foreign Language Speaking Proficiency

Special Assessment Techniques

- Face-to-face interview of 20-25 minute duration
- Student proficiency evaluated in terms of performance capabilities in real-life situations
- Rating scale encompasses entire range of speaking proficiency from "survival" level to proficiency shown by educated native speakers

Comments. This language proficiency interview is based on techniques originally developed by the U.S. Foreign Service Institute and recently adapted for use in Peace Corps language training, teacher certification programs, and other situations. The interview is conducted by either one or two trained raters who converse with the examinee in a relatively informal setting on topics representative of a number of real-life communication areas. Throughout the interview, the pace and level of sophistication of the discussion are increased to a point at which the maximum level of performance has been attained. An 11-point scale (5 basic performance levels, together with intermediate "plus" values) is used to rate the total performance. Each of these levels is defined by a short paragraph describing the kinds of language-use situations in which the examinee is considered capable of communicating in an effective and appropriate manner.

Concluding Note

Several of the above illustrations contain elements of product assessment, namely numbers 3, 7, and 8. However, it should be stressed that the principal focus of the assessment is upon the performance of the student in a situation where the product is introduced as a means of identifying the performance skills necessary to create the product.

Bibliography On Performance Assessment

Boyd, Joseph L., Jr., and Shimberg, Benjamin. *Handbook of performance testing: A practical guide for test makers*. Princeton, N.J.: Educational Testing Service, 1971, 182 pp.

Breen, Paul; Donlon, Thomas F.; and Whitaker, Urban. *Learning and assessing interpersonal competence–A CAEL student guide*. Princeton, N.J.: CAEL, Educational Testing Service, 1977.

Breen, Paul; Donlon, Thomas F.; and Whitaker, Urban. *Teaching and assessing interpersonal competence–A CAEL handbook*. Princeton, N.J.: CAEL, Educational Testing Service, 1977.

Colwell, Richard, et al. Tests and reviews: Fine arts-music, in Oscar K. Buros (Ed.), *The seventh mental measurements yearbook*. Highland Park, N.J.: Gryphon Press, 1972, vol. 1, pp. 526-536.

Crooks, Lois A. *Issues in the development and validation of in-basket exercises for specific objectives*. Research Memorandum 68-23. Princeton, N.J.: Educational Testing Service, 1968, 13 pp.

Fitzpatrick, Robert, and Morrison, Edward J. Performance and product evaluation, in Robert L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971, pp. 237-270.

Forrest, Aubrey. *Assessing prior learning–A CAEL student guide*. Princeton, N.J.: CAEL, Educational Testing Service, 1977.

Franks, B. Don, and Deutsch, Helga. *Evaluating performance in physical education*. New York: Academic Press, 1973, 226 pp.

Frederiksen, Norman. Factors in in-basket performance. *Psychological Monographs: General and Applied*, 1962, vol. 76, no. 22, pp. 1-25 (Whole No. 541).

Gagne, Robert M. Simulators, in Robert Glaser (Ed.), *Training research and education*. Pittsburgh, Penn.: University of Pittsburgh Press, 1962, pp. 223-246.

Kelley, Edgar. *Three views of competency-based teacher education. III. University of Nebraska*. Bloomington, Ind.: Phi Delta Kappa Educational Foundation, 1974, 29 pp.

Knapp, Joan. *Assessing prior learning–A CAEL handbook*. Princeton, N.J.: CAEL, Educational Testing Service, 1977.

Knapp, Joan, and Sharon, Amiel. *A compendium of assessment techniques*. Princeton, N.J.: CAEL, Educational Testing Service, 1975, 50 pp.

Lehman, Paul R. *Tests and measurements in music*. Englewood Cliffs, N.J.: Prentice-Hall, 1968, 99 pp.

Lopez, Felix M. Measuring human performance. Chapter 8 in Felix M. Lopez, *Evaluating employee performance*. Chicago, Ill.: Public Personnel Association, 1968, pp. 163-185.

McCormick, Ernest J., and Tiffin, Joseph. Performance evaluation, in Ernest J. McCormick and Joseph Tiffin, *Industrial psychology* (6th ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1974, pp. 193-219.

McGee, Rosemary and Drews, Fred. *Proficiency testing for physical education*. Washington, D.C.: American Association for Health, Physical Education, and Recreation, 1974, 73 pp.

McGuire, Christine H., and Babbott, David. Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement*, 1967, vol. 4, no. 1, pp. 1-10.

Meyer, Herbert H. The validity of the in-basket test as a measurement of managerial performance. *Personnel Psychology*, 1970, vol. 23, pp. 297-307.

Rivas, Frank W. The first national assessment of musical performance, in *National assessment of educational objectives*. (Report 03-MU-01) Denver, Colo.: Education Commission of the States, 1974, 29 pp.

Sharon, Amiel T. *Planning the development of measurement and evaluation services for use in occupational programs at postsecondary institutions.* (PR 74-16) Princeton, N.J.: Educational Testing Service, 1974. 54 pp.

Watkins, John G. *Objective measurement of instrumental performance.* New York: Teachers College, Columbia University, 1942. 88 pp.

VI. Assessment of Written Material

Myrna Miller

The outcomes of learning that has been gained through experience in a nonclassroom setting are most often communicated via the written mode, and experts are called on to measure and evaluate this written work. The form of the work is usually one of the following: an essay examination answer, an assigned topic essay, a general essay within the portfolio of prior experience, or a log, diary, or journal. In addition to the general concepts of reliability and validity, experts making academic judgments about written materials need to pay particular attention to the pitfalls implicit in three significant questions: (1) Is the written work a *direct* sample of the learning claimed? (2) Is the written work an *assertion* that learning has occurred in the past or in a distant setting? (3) Is the written work an indirect sample from which the expert can deduce cognitive and affective learning?

Of these three types, the *direct* sample is most like the product of classroom learning. In general the assessment of this type of written material requires little modification for application to experientially acquired learning. Because of the similarities to other product assessment, be sure to consult Chapter IV for the discussion of important issues in product assessment.

The direct sample written product is often an essay examination answer, an assigned topic essay, or research report. If the student claims to have learned American history through travel or self-directed reading, he can be given an essay examination by an expert historian. The historian can measure and evaluate the student's knowledge by using the appropriate criteria for what constitutes a correct response. In another case, if the student seeks credit for English composition or communication skill competence, an assigned topic essay can be evaluated as direct evidence of written verbal ability and knowledge.[1] A third direct sample might be a research paper submitted in Spanish for a student claiming to have learned Spanish while living abroad. Such direct samples are clearly useful for a competence currently held and easily demonstrated through summative evaluation.

All direct sample written responses require explicit agreed-upon criteria for evaluation, which have been made available to both the assessor and the student before the actual evaluation occurs. In the classroom each student gradually and subtly comes to know the professor's expectations. Although the professor may not have made known his standards at the outset of class, experienced students usually learn to infer the criteria for evaluation and modify their study plans accordingly. This ability to discern evaluative criteria may, in fact, be the actual measured skill of "successful" students.

Continued classroom interaction throughout the semester gives classroom students the opportunity to sort out information relevant to expectations for measurement and evaluation. In addition, the experience with mid-course formative assessment processes prepares students through trial-and-error learning for final

---

[1]For a discussion of the topical essay as a technique for assessing interpersonal skills, see Paul Breen, Thomas F. Donlon, and Urban Whitaker, *Teaching and Assessing Interpersonal Competence--A CAEL Handbook* and *Learning and Assessing Interpersonal Competence--A CAEL Student Guide*.

summative evaluation. Usually the student whose learning has been experientially derived cannot reap these benefits of prior knowledge of the professor or formative assessment. Generally, the learning outcomes of experience are judged by an unfamiliar expert. Whenever this is the case, every effort should be made to clarify what is expected, to make known what criteria for evaluation are to be applied, and to describe all other "rules" for assessment.

Students in on-going experiential settings often keep logs, diaries, or journals of their daily learning activities. These written responses are then submitted to experts for assessment of learning. Throughout such journals, there are assertions that learning is taking place such as, "I am finally beginning to enjoy typing now that I am only correcting one error per purchase order." Acceptance of such an assertion should not be substituted for an actual typing test or a letter of verification from a supervisor. This rather obvious example is given in order to contrast it with the more common and more difficult cases that frequently arise. For example, a student working on a psychiatric ward may claim to be learning abnormal psychology through his daily observation of patients. Another example might be when a student working in a nursery school describes the children's daily activities and claims to be learning early childhood development. In these cases, the task of the expert is to distinguish between those statements that are mere reporting of observations and those that indicate thoughtful analyses based on college-level conceptual frameworks.

This issue of assertion versus true learning outcomes is crucial in the assessment of experiential learning. The distinction is, of course, often extremely difficult to make. Most of us are unwittingly committed to the belief that seeing is learning (on some level). Our children may say, "I learned a lot today. I went to the fire station and saw the fire engines," or a friend may report, "I went to India and saw the Taj Mahal." In truth, these stimulating visual experiences plus the personal experiences accompanying them, so often do lead to learning that we make the incorrect assumption that all such experiences lead to significant learning for all learners. *Experts need to keep in mind that the experience, setting, or activity is not the learning; it is the opportunity for learning to occur. Therefore, in making judgments, assessors must carefully sort out the student's descriptions of learning opportunities from actual college-level learning deserving the award of credit.*

The portfolio of prior learning is especially susceptible to this problem of distinguishing opportunity from learning. It is difficult for even the most experienced assessor to sort out actual learning from very persuasive descriptions of learning. Talented students with exceptional writing ability can often provide very convincing cases for credit. The expert may be crediting writing expertise when called on to evaluate other skills. Or he or she may assume a connection between the writing ability and the learning outcome claimed (see "halo effect," Chapter II). It is often even more difficult for the student to sort out his/her own learning from experience, and portfolio preparation almost always requires that students disentangle learning from experience.[2] Most programs have found that a strong counseling component is particularly necessary for older students who have been away from the academic world for some time.

---

[2]For detailed discussions of portfolio preparation and assessment, see Joan Knapp, *Assessing Prior Learning--A CAEL Handbook*, and Aubrey Forrest, *Assessing Prior Learning--A CAEL Student Guide*.

Because past learning is a unique personal experience, many experts may feel that student assertions must be accepted. After all, the student is the verifier that he or she was there at the time. This does not mean, however, that the students' assertions are synonymous with learning outcomes. Again a good counseling system can help students to develop self-assessment skills that can provide reliable statements of learning outcomes. In this way, not only the faculty assessor benefits by having clear statements of learning outcomes as the foci of assessment, but the students also gain a new usable skill.[3]

Although it is essential to separate mere assertion from actual learning outcomes, the separation of personal experience from individually internalized learning is somewhat artificial. Human lives, experience, feelings, thoughts and learning, are in fact, all of one piece. Students oftentimes respond emotionally to judgments of past learning activities and frequently perceive these judgments as evaluations of the worth of their lives. According to Richard G. Beery, in his article "Fear of Failure in the Student Experience," our society reinforces the idea that the measurement of achievement is equal to the measurement of ability, which is interpreted by many as the worth of one's life. For this reason, it is highly recommended that experts evaluating prior learning always be disinterested parties free from constraints in distinguishing learning from described activities that might be construed as the total life's worth.

The written response, whether essay exam, assigned essay, portfolio essay, or journal, can often provide a sample of the student's cognitive and (less frequently) affective processes from which an expert can deduce creditable learning outcomes. If we select from Bloom's Taxonomy,[4] we can choose specific cognitive processes to be used as criteria for judging written materials. An example might be a student who was a Peace Corps volunteer and who describes his experience with the problem of offsetting malnutrition in a remote South American village. He might indicate a comprehensive *knowledge* of the situation by providing essential data on eating habits, agricultural practices, birth rates, death rates, etc. He might demonstrate his ability to *analyze* the problem through discussion of the factors he chose to assign to higher and lower priority. *Synthesis* might have been demonstrated through the combination of resources he used to address the problem. The ability to *evaluate* might have been shown in his choice of priority to attack, i.e., encouraged change in eating habits when lack of new equipment made a change in agricultural practices unfeasible.

In a similar manner, interpersonal skills and understanding of others might be assessed by reading this student's description of his personal adjustment to the mores, customs, and role expectations of this new culture. The expert should not base his judgments on a statement such as "I grew and changed a lot," or "I really learned to understand the people I worked with." The evaluation, although based on the student's descriptions of his participation in the delivery of a baby, should be derived from thoughtful reflection on this experience. If he was able to provide "help" within the context of a second culture, and can provide details of his self-

---

[3]For a discussion of the student as expert, see "Assessors and Their Qualifications" by Urban G. Whitaker, in Morris T. Keeton and Associates, *Experiential Learning: Rationale, Characteristics, and Assessment*.

[4]Benjamin S. Bloom (Ed.), *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: The Cognitive Domain*.

awareness and understanding through discussion of his personal attitudes (culturally derived) toward the mother, the methods of delivery, and the infant, then credit for these skills may be quite appropriate. When attempting judgments of cognitive and affective processes, experts must be careful to avoid the pitfalls of crediting assertions, and, in this case, not award credit for the "biology of child-birth," which has not, in fact, been demonstrated. Or, if our Peace Corps volunteer happens to have outstanding writing ability, the experts must avoid the pitfall of crediting verbal facility when called on to judge analytic, evaluative, or interpersonal ability.

CATEGORIES OF WRITTEN MATERIAL OFTEN CONSIDERED FOR CREDIT

Essays

Much of the research literature has been devoted to contrasting essay examinations with objective examinations. In assessing experiential learning, the basic assumption is that objective exams would be too costly to produce for the limited number of students who have had the same or nearly similar experiential learning. Those students who have acquired either broad learning or highly specific learning similar to traditional course work can take the already available CPEP or CLEP exams. The concern of this *Handbook* is with the assessment of unique experiential learning both in sponsored settings and prior learning. For these purposes the essay may be appropriate.

Experts who are asked to compose essay questions or to rate essay answers should be provided a common definition of this type of assessment task. A definition which includes the weaknesses of the essay and which indicates possible problematic areas is provided by Stalnaker.

> . . . the essay question is defined as a test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct, and the accuracy and quality of which can be judged subjectively only by one skilled or informed in the subject. The most significant features of the essay question are the freedom allowed the examinee and the fact that not only can no single answer be listed as correct and complete, and given to clerks to check, but even an expert cannot usually classify a response as categorically right or wrong. Rather, there are different degrees of quality or merit which can be recognized.[5]

When experts are expected to make judgments on degree of quality, they must be provided with clear criteria. Most writers suggest that essays should be read quickly for an over-all impression, especially when they are being judged for quality of writing as well as for substance and when there is a relatively large sample of papers on the same topic. Readers should be asked *not* to analyze specific points or to note trivial weaknesses. A holistic or global approach is generally favored. Even if there is only one essay on a given topic, it should be read by at least two judges whenever possible.

---

[5] J. M. Stalnaker, "The Essay Type of Examination," p. 495.

The tasks should be clearly defined so that the student can be made aware of what is expected. All questions should be carefully worded so that the student understands what he is expected to do. If an answer of a particular length is expected, specify how long it is to be. If the student is allowed to expand upon the question, state this. If he is expected to stay within given limits, make the boundaries known to him. When the students are expected to set aside mechanics and wording and to concentrate on ideas, form, and flavor, they should be aware of exactly what they are being graded on, and readers should be selected who agree to grade these factors. Sometimes, of course, readers who agree to grade on these factors may nevertheless find themselves influenced from time to time by grammar, spelling, etc. Students therefore should be forewarned that it may be impossible for evaluators to totally ignore mechanics and wording and that errors in grammar, spelling, etc., may affect the evaluation.

Portfolios

The portfolio of prior learning usually contains as one portion a general essay. This essay is frequently assessed by a faculty expert or panel of experts. The essay is often autobiographical and highly descriptive of past experiences. The essays can range from a few pages to full booksize manuscripts. Details may include emotional reactions to past events and intrapersonal insights as well as job descriptions and catalogs of skills attained. With different portfolios, the experts who assess the learning in the portfolio may do so from different perspectives and with different criteria for assessment. For any one portfolio, however, the criteria should remain the same across experts.

1. Descriptions of Past Learning as a Measurement of Writing Competence. Most institutions of higher education have as a stated requirement for the degree an acceptable level of writing competence. This competence is frequently met through the preparation of the portfolio essay. Experts who are called on to assess the portfolio essay should be provided with a single set of criteria specific to this purpose such as:

a. Clarity of presentation, proper organization, appropriateness of style.
b. Logic and order of presentation, adequacy of transitions, unity, and coherence.
c. Neatness, lack of typographical or spelling errors.
d. Accuracy of punctuation.
e. Correctness of grammar, adequacy of vocabulary.

The faculty of each institution needs to define and establish criteria appropriate to the student body and consonant with the philosophy and standards of that institution.

2. Articulation with Degree Program. The portfolio essay frequently serves as the vehicle of articulation of past learning with learning to be undertaken at the institution. The task, in this case, is to assess the student's ability to analyze the significant, relevant learning from the past and to integrate this learning into a full

degree program. Depending on the particular objectives or goals of the institution, the criteria for this assessment will differ. Experts will require a set of clear criteria when they are called upon to evaluate the complex processes involved in dredging up, sorting out, categorizing, and articulating past learning. An example might be a student who decides to study oceanography after many years of working as an engineer. His work experience might provide him with the knowledge of geophysics and other course work he needs. Much of his earlier academic work may now be rusty or out-of-date and may need to be discarded. Some of his Navy training may be, after close examination, highly relevant. Also his avocations of sailing, boatbuilding, and marine natural history may fit in well with a study plan that calls for an at-sea internship.

Criteria for the assessment of this type of essay might include:

a. Ability to present a well-organized logical argument.
b. Ability to identify pertinent learning from experience.
c. Ability to make decisions with well-supported defenses.
d. Ability to justify and to substantiate each decision.

Logs, Diaries, Journals

In the traditional classroom, the faculty member responsible normally has control over the material presented, the quality of work to be undertaken, and the quality of the readings or other materials to be used. Since the subject matter emanates from the instructor in the form of assignments or lectures, there is little doubt about what was taught—only about what was learned. This controlled teaching process allows each faculty member to devise measurement instruments based on what was delivered and, hopefully, received. In the experiential setting, however, much of this teacher control is given over to nonacademic supervisors, employers, or the student. As a consequence, one rather common method of recording the learning activities has been the log, diary, or journal. The student submits this log to an expert for evaluation either as interim or final documentation of learning. When a log is used without on-site observation, the expert is called on to evaluate both what was delivered (available learning experience) and what was received (learning outcomes).

1. Description of Learning Activities. Often the log, diary, or journal takes the form of a direct daily record of each learning activity. Students in internships, on-the-job training, or apprenticeships may include a fine level of detail for each new task undertaken, level of responsibility accepted, or skill developed. An example might be a teacher's aide who is assigned one child with a reading problem to work with and describes for each day the phonic drills, word games, and stories used as part of the student's teacher-as-learner experience. An expert reading such a journal could assess the learning by using criteria such as:

a. Ability to apply theory to actual situation.
b. Ability to interrelate facts, principles, and phenomena in a new way.
c. Ability to demonstrate the development of a new set of interrelated concepts.

2. Reflection, Cognition, Affect. The diary or journal can also be used to record the students' thoughts about and feelings toward their experiential learning. It is often this very reflection upon the experience that produces the true learning outcomes. A student of drama may gain more insight into her art after considerable reflection than on-stage during the actual performance. A social work intern, after many hours of informal self-analysis, may recognize that she has been too quick to make decisions for her clients. Students of writing and literature may use the journal to record books read and reactions to the books, as well as personal creative efforts such as poems or short stories.

Criteria for the reflective journal might include:

a. Ability to draw analogies between related experiences.
b. Ability to integrate experience into larger patterns of meaning.
c. Ability to demonstrate that the experience shaped the student's views or philosophy of life.

(See CAEL Institutional Report No. 3, University of California at Los Angeles, for further discussion of the use of the journal for sponsored experiential learning).[6]

GENERAL CONSIDERATIONS

Important steps that should be taken into account when assessing written materials include the following:

1. Determine the specific factors on which the student is to be assessed before making any judgments. Make sure that the *student* has a clear understanding of exactly what factors (e.g., writing competence, substantive knowledge in some area, etc.) are to be assessed.
2. Obtain a representative sample of the student's writing.
   a. If the assessment is to be based on prior written work, provide guidelines to the student so that a representative sample of writing related to the criteria can be assessed. If a student claims to have written 20 short stories, for example, the assessment should be based on more than one or two of the stories.
   b. If the assessment is to be based on a requested writing sample, the directions and guidelines to the student should be clear and relevant to the factors established.
3. Prepare for assessment.
   a. Through committee action or other methods, standards should be established. For example, a file containing samples of acceptable levels of student work can be set up.
   b. Clear guidelines should be established for judges so that they know exactly what factors are to be assessed.
   c. Through workshops and/or practice sessions, judges should be trained to use agreed-upon procedures and standards.

---

[6]Jane Szutu Permaul and Marina Buhler Miko, *Documentation and Evaluation of Sponsored Experiential Learning*.

4. Make final assessment report.
   a. Clear guidelines should also be provided for the writing of a final assessment report. If the report is to be of diagnostic value, for example, the assessors should have some predefined structure around which such a report can be written. The report should clearly relate the final assessment to the written material.
   b. Whenever possible have at least *two* judges read the material. Each judge should not know the other's evaluation prior to doing his or her assessment.
   c. As a rule, judges who are personally familiar with the writer should not be used. Whenever possible, the student author of the materials should remain unidentified.

Bibliography on Assessment of Written Material

Beery, Richard G. Fear of failure in the student experience. *Personnel and Guidance Journal*, 1975, vol. 54, no. 4, pp. 200-201.

Bloom, Benjamin S. (Ed.) *Taxonomy of educational objectives: The classification of educational goals.* Handbook I: *Cognitive domain.* New York: David McKay, 1956.

Coffman, William E. On the reliability of ratings of essay examinations. *NCME Measurement in Education*, 1972, vol. 3, no. 3. 7 pp.

Coffman, William E. Essay examinations, Chapter 10 in Robert L. Thorndike (Ed.), *Educational measurment* (2nd ed.). Washington, D.C.: American Council on Education, 1971, pp. 271-302.

Ebel, Robert E. The characteristics and uses of essay tests, in Robert E. Ebel, *Essentials of educational measurement* (2nd ed. of book formerly titled *Measuring educational achievement*). Englewood Cliffs, N.J.: Prentice-Hall, 1972, pp. 123-154.

Forrest, Aubrey. *Assessing prior learning--A CAEL student guide.* Princeton, N.J.: CAEL, Educational Testing Service, 1977.

French, John W. Schools of thought in judging excellence of English themes, in Educational Testing Service, *Proceedings of the 1961 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1961, pp. 19-28.

Godshalk, Fred I., Swineford, Frances, and Coffman, William E. *The measurement of writing ability.* Research Monograph No. 6. New York: College Entrance Examination Board, 1966, 88 pp.

Knapp, Joan. *Assessing prior learning--A CAEL handbook.* Princeton, N.J.: CAEL, Educational Testing Service, 1977.

Nealey, Stanley M. Student-instructor agreement in scoring an essay examination. *Journal of Educational Research*, 1969, vol. 3, no. 3, pp. 111-115.

Permaul, Jane Szutu, and Miko, Marina Bühler. *Documentation and evaluation of sponsored experiential learning.* CAEL Institutional Report No. 3, University of California at Los Angeles (revised ed.). Princeton, N.J.: CAEL, Educational Testing Service, 1977.

Stalnaker, John M. The essay type of examination, in E.F. Lindquist (Ed.), *Educational measurement.* Washington, D.C.: American Council on Education, 1951, pp. 494-530.

Whitaker, Urban G. Assessors and their qualifications, Chapter 12 in Morris T. Keeton and Associates, *Experiential learning: Rationale, characteristics, and assessment.* San Francisco: Jossey-Bass, 1976, pp. 189-223.

A CONCLUDING NOTE

Richard R. Reilly

The preceding chapters have each presented some important considerations for assessment. The six basic steps which were recommended in Chapter II reoccur, with differing emphases on various points, in the chapters dealing with specific methods of assessment. It may be helpful to review these steps as they apply to each type of assessment. Figure 5, "A Matrix of Assessment Steps," presents in a concise way the specific adaptation of each general step as it pertains to each of the four major assessment techniques discussed in this *Handbook*.

Figure 5 clearly illustrates the similarity in basic principles that should be applied to any assessment. The first major step, *Establish Criterion Standards*, is a good example. The idea is essentially the same for all four types of assessment methods. Concrete examples or descriptions should be used to illustrate what is meant by different levels of the criterion. If performance is being assessed, this might be done with a number of direct examples of performance. If products are being assessed, actual examples of products would be best, or if that is not possible, clear verbal descriptions of products could be used. The second basic step includes two activities. First, within each of the four assessment techniques, there is a range of possible approaches. Careful consideration of the assessment problem at hand will usually reveal one or two approaches best suited to the situation. If interviewing is used, this may mean choosing an oral test rather than a semistructured interview, for example. The second activity in this step involves providing some structure for the assessee and assessor. Structure is what distinguishes an interview from a casual conversation at a cocktail party or product assessment from a visit to an art show.

The next step, *Planning*, includes both the preparation of judges and the administrative details, such as providing an adequate room. An environment free of distracting noise is clearly essential for effective interviewing, and for many types of performance assessments special equipment will be needed.

The fourth step has to do with directing and focusing the assessor's attention to only what is observed within the structure of the assessment procedure. The age or appearance of the artist should not in any way influence the judgment of assessors considering artistic products as learning outcomes, for instance. Once judgements have been made, it is important that documentation of the assessment process be available for review and that feedback be given to the student. This fifth step may vary considerably in terms of the mode of transcription. Products, for example, serve as a record themselves, whereas interviews must be electronically recorded or detailed with written notes.

The sixth and final step can take a variety of forms, ranging from judgmentally evaluating the adequacy of an assessment procedure to a full-blown empirical research study. Monitoring the quality of assessment is a crucial, but often neglected, activity. Within a given institution it is often difficult or impossible to apply the more powerful empirical procedures to evaluating assessment simply because not enough students are assessed with a given procedure. Empirical procedures

Figure 5
A Matrix of Assessment Steps

Basic Steps	*Assessment Techniques*			
	Interviews	*Product Assessment*	*Performance Assessment*	*Written Material*
Establish Criterion Standards	Should be given in behavioral and verbal terms.	Should be defined by specific examples.	Should be given in behavioral terms.	Should be related to objective, observable aspects of writing.
Select and Structure Assessment	Determine what type of interview is to be conducted and structure it to elicit relevant information.	Provide guidelines for choosing the most relevant and representative products.	Choose the most relevant performance method and adapt to specific criteria.	For requested written material, specific guides should be provid for the student. For previously produced material, guidelines should be given for the number amount of material to be submitted.
Plan the Administration of Assessment	Prepare interviewers, interview schedules, provide for space, time, and special recording equipment.	Prepare judges by providing them with clear specifications and examples.	Prepare judges, provide for space, time, and any special equipment needed.	Prepare judges by providing th with clear instructions as to wh they will be judging.
Relate Judgments to Observations	Use of checklists, rating scales, and written reports.	Rating scales and written reports.	Checklists, rating scales, and written reports.	Rating scales and written repo
Record and Report	Recording can be electronic and written. Feedback given in postinterview.	Products themselves serve as a record. Also written records. Feedback written or oral.	Recording can be electronic *and* written. Feedback can be written or oral.	Recording is written only. Feedback can be given in writ or oral form.
Monitor Assessment	Tape recordings or additional observer can be used, in addition to examination of written reports and empirical study.	Informal inter-judge agreement, written reports, and empirical study.	Internal consistency of the tasks and electronic recording are some possibilities, in addition to written reports and empirical study.	Informal inter-judge agreement examination of written reports, and empirical study.

require relatively large samples for results to be meaningful, and it is unfortunately true of most CAEL institutions that the required sample sizes cannot normally be obtained. Special studies that draw on the resources of several institutions can overcome the "numbers" problem and provide some valid answers to questions about assessment quality. In order to do this, however, common procedures must be applied in different institutions, faculty cooperation must be obtained, and a host of difficult problems must be confronted and overcome. The difficulty of such an undertaking is what makes the CAEL validation studies[1] such a unique and meaningful accomplishment. These studies drew on the cooperative efforts of 24 institutions, all involved in assessment of nontraditional learning. The results of these studies provide, for the first time, "base rate" information on the reliability and validity of assessment of experiential learning and, as such, would be an ideal starting point for anyone involved in evaluation of assessment at the local level.

---

[1]Warren W. Willingham and Associates, *The CAEL Validation Report*. Princeton, N.J.: CAEL, Educational Testing Service, 1976.

CAEL HANDBOOKS AND CAEL STUDENT GUIDES

Teaching and Assessing Interpersonal Competence—A CAEL Handbook	$6.00
Assessing Prior Learning—A CAEL Handbook	$6.00
College-Sponsored Experiential Learning—A CAEL Handbook	$6.00
Expert Assessment of Experiential Learning—A CAEL Handbook	$6.00
Assessing Occupational Competences—A CAEL Handbook	$6.00
Learning and Assessing Interpersonal Competence—A CAEL Student Guide	$3.00
Assessing Prior Learning—A CAEL Student Guide	$3.00
College-Sponsored Experiential Learning—A CAEL Student Guide	$3.00

A list of other CAEL publications is available upon request.